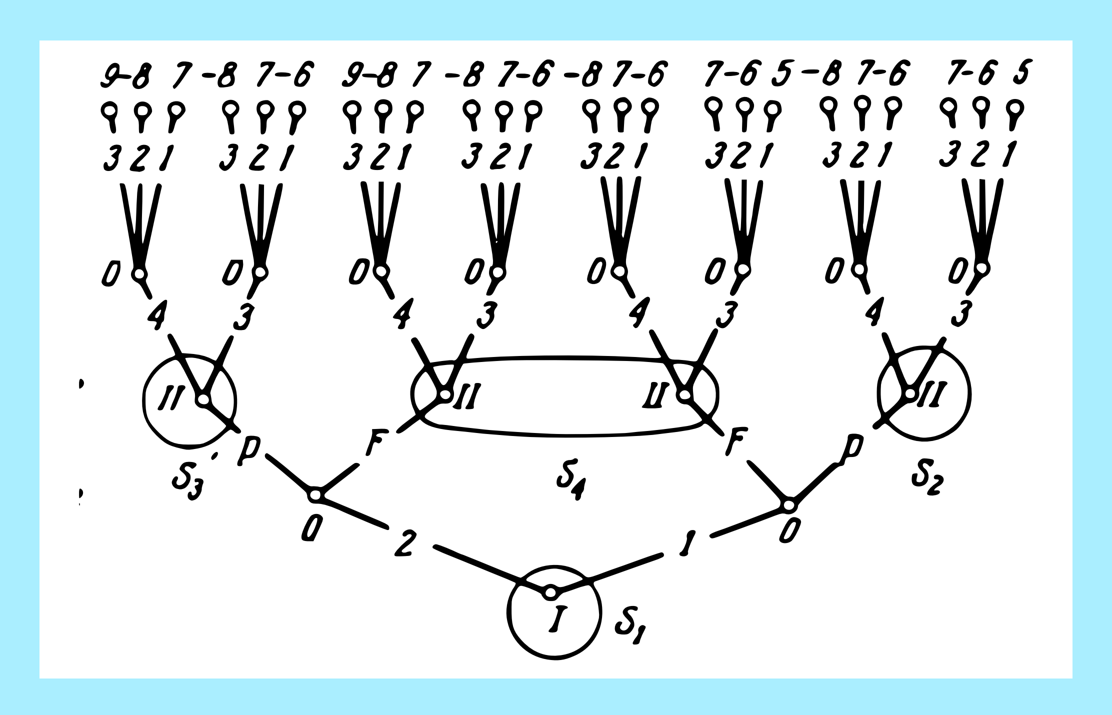
Y. Korchounov

FONDEMENTS MATHÉMATIQUES DE LA CYBERNÉTIQUE



Éditions Mir Moscou

Ю. М. КОРШУНОВ

МАТЕМАТИЧЕСКИЕ ОСНОВЫ КИБЕРНЕТИКИ

ИЗДАТЕЛЬСТВО «ЭНЕРГИЯ»
МОСКВА

Y. KORCHOUNOV

FONDEMENTS MATHÉMATIQUES DE LA CYBERNÉTIQUE

ÉDITIONS MIR · MOSCOU

Traduit du russe

par Vladimir Kotliar

На французском языке

AVANT-PROPOS

Il est d'usage de désigner sous le terme « cybernétique » tout un essaim de nouvelles théories scientifiques, qui, apparues au cours des dernières décennies, trouvent un emploi remarquablement large dans le domaine de l'automatique et de la télémécanique. Les cours professés aux étudiants qui se consacrent à ces spécialités, comportent des notions de la théorie de l'information, de la théorie de la détection statistique des signaux, de la théorie des systèmes optimaux. de la théorie des automates finis, ainsi qu'une série de principes relevant de la théorie des jeux et de la théorie des décisions statistiques.

Or, le professeur qui se propose d'enseigner aux étudiants des problèmes de la cybernétique se heurte à une difficulté sérieuse: le cours traditionnel des mathématiques supérieures lu dans les écoles supérieures d'enseignement technique est axé sur le *continu* et le *déterminé*, ce qui est insuffisant pour l'étude des nouvelles disciplines, basées principalement sur le *discret* et l'*aléatoire*. Ces questions peuvent être traitées soit dans le Cours des fondements mathématiques de la cybernétique, soit dans le Cours des fondements théoriques de la cybernétique.

L'ouvrage proposé est basé sur les conférences lues aux étudiants de la spécialité « Automatique et Télémécanique » et de quelques spécialités connexes à l'Institut Radiotechnique de la ville de Riazan.

Le présent Cours comprend une Introduction et deux parties divisées en dix chapitres.

A l'Introduction, à côté de la notion de système cybernétique, on donne au lecteur certains renseignements sur les systèmes de communication et de commande, afin de pouvoir illustrer ensuite les méthodes mathématiques exposées par des exemples familiers aux spécialistes dans ce domaine.

La première partie contient plusieurs chapitres de mathématiques absolument indispensables à l'ingénieur. A elle seule, la première partie est déjà d'une grande utilité, car les méthodes qu'elle expose permettent de résoudre toute une série de problèmes pratiques importants du domaine de l'automatique et de télémécanique. Par ailleurs, les matières y sont disposées de telle façon que la première partie constitue une sorte d'initiation mathématique à la deuxième partie

qui est consacrée à l'optimisation des processus de commande et qui se donne pour tâche d'initier l'étudiant aux notions de la théorie moderne de la gestion.

Dans ce petit nombre de pages, bien évidemment, nous n'avons pu exposer qu'un petit nombre de notions et méthodes de la cybernétique; cela est vrai tant pour la première partie que (surtout) pour la deuxième. En particulier, l'optimisation des processus de commande, qui représente un chapitre essentiel de la théorie générale de la gestion, fait un large usage des procédés de l'analyse mathématique classique et du calcul variationnel classique. Les méthodes d'optimisation de la commande, basées sur ces procédés et traitées généralement dans le Cours de la théorie de la commande automatique et dans plusieurs cours plus spécialisés, n'ont pas trouvé de place dans ces pages. Par contre, nous avons accordé le maximum d'attention aux idées et méthodes qui ont vu le jour à cause de l'évolution de la cybernétique et des ordinateurs. Or, même dans ces cas, on a été souvent obligé de ne faire qu'une description assez sommaire des idées fondamentales.

Il est à noter que, bien que la presque totalité des méthodes exposées ait été illustrée par des exemples et des problèmes à résoudre, c'est encore la résolution de problèmes par ses propres forces qui reste pour l'étudiant l'unique moyen d'apprendre à fond les méthodes de calcul numérique; aussi ce cours devra-t-il être complété par des cours pratiques et par des devoirs de calcul distribués aux étudiants de la spécialité.

Nous tenons à exprimer notre profonde reconnaissance à L. Kouzine et D. Pospélov, qui ont bien voulu se charger de la critique du livre, et au rédacteur N. Glazounov, dont les remarques très précieuses ont beaucoup ajouté à la valeur du livre.

Y. Korchounov

INTRODUCTION

I-1. OBJET DE LA CYBERNÉTIQUE

Science jeune, apparue au lendemain de la Seconde guerre mondiale, la cybernétique a connu un essor tellement prodigieux qu'elle occupe à présent, dans de nombreux domaines de la science et de la technique, des positions très stables. Les succès de la cybernétique sont dus à la découverte d'un certain nombre d'analogies existant entre le fonctionnement des dispositifs techniques, la vie des organismes et le comportement des collectivités d'êtres vivants. La cybernétique a complété ces analogies, qui résultent de certains raisonnements généraux de caractère méthodologique, par des méthodes mathématiques spéciales permettant de donner une définition quantitative aux processus se produisant dans des systèmes physiques très variés. Les principes de la cybernétique trouvent un large emploi en automatique, télémécanique, théorie des communications, économie, sociologie, biologie et médecine [1 à 5].

Le mot « cybernétique » vient de la langue grecque. Les Anciens employaient ce mot pour désigner l'art de gouverner le navire. Au XVIIIe siècle nous rencontrons ce mot dans les ouvrages de A. M. Ampère, grand physicien et mathématicien français, qui définit par ce mot la partie de la politique qui s'occupe des moyens de gouverner. Nous entendons aujourd'hui, sous ce terme, la science de la gestion au sens très large. Le contenu moderne de cette notion est lié au nom du savant américain Norbert Wiener, dont le livre « Cybernetics, or Control and Communication in the Man and the Machine » paru en 1948 a posé la première pierre de cette discipline moderne.

L'apparition de la cybernétique en tant que science de la gestion est étroitement liée au progrès technique général caractérisant l'évolution des forces de la production à l'époque actuelle.

Avant l'apparition de la cybernétique, les techniciens voyaient leur tâche principale, premièrement, à mettre au point les dispositifs servant à obtenir et à transformer l'énergie (machines à vapeur, turbines, génératrices électriques, moteurs électriques, etc.), et deuxièmement, à créer les dispositifs susceptibles d'agir sur l'environnement. De pareils dispositifs se caractérisaient essentiellement par des rapports énergétiques, et leur paramètre principal est le

rendement, ou le débit. Assez élémentaires du point de vue technique, ces dispositifs ne posaient aucun problème quant à leur commande. L'homme pouvait travailler et commander l'objet de son travail en même temps. L'information nécessaire pour la commande lui parvenait directement par ses sens quand il observait les résultats de ses efforts.

Cependant, au milieu du XXe siècle, le progrès de la science a permis la construction de systèmes techniques tellement compliqués que leur commande représentait pour l'homme une tâche physiologiquement impossible. Par exemple, vers la fin de la Seconde guerre mondiale, les chercheurs se sont trouvés devant le problème de création d'un système de conduite de tir automatique antiaérien tel qu'il puisse, sans la participation de l'homme, assurer la poursuite des avions volant à des vitesses comparables à celles des projectiles antiaériens, effectuer le calcul de leur trajectoire et réaliser le pointage des canons. Dans de pareils systèmes, il s'agit en premier lieu d'obtenir l'information sur la situation du moment, de traiter cette information de manière à en tirer les données utilisables pour la commande et de mettre à profit cette information pour des actions ayant un but précis; en d'autres mots, le problème se ramène à créer certains dispositifs *de communication* et *de commande.* La nécessité de résoudre ces problèmes a stimulé les rapides progrès réalisés dans la théorie des communications, dans les techniques de calcul et dans l'automatique; les notions dégagées au cours de ces travaux furent ensuite mises à la base de la cybernétique.

La différence fondamentale entre les dispositifs de communication et de commande d'une part et les dispositifs techniques dont il a été question plus haut d'autre part réside dans le fait que les rapports énergétiques cessent désormais de jouer le rôle prépondérant; par contre, l'attention est accordée principalement à la faculté de ces dispositifs de transmettre et de traiter de grandes quantités d'information sans les déformer en aucune façon. C'est ainsi que dans la liaison radio le récepteur ne capte qu'une partie infime de l'énergie émise par l'antenne de l'émetteur, de sorte que le rendement se trouve très bas. Or, la bonne qualité d'une ligne de transmission consiste en ce que les messages captés ne subissent que de très petites déformations et que l'influence des parasites est exclue. Donc, les processus fondamentaux se produisant dans les systèmes de communication et de commande sont les processus de transmission et de traitement de l'information, et non pas les processus liés à la transformation et à l'utilisation de l'énergie.

L'insignifiance des rapports énergétiques dans les problèmes de communication et de commande permet de faire abstraction des particularités physiques des supports d'information et de la nature physique des systèmes dans lesquels cette information est utilisée; aussi la cybernétique est-elle appelée à fournir une théorie générale

de la communication et de la commande, applicable à tous les systèmes, quelle que soit leur nature physique.

La notion de *système*, à côté de celle de *gestion* (ou de *commande*), est la notion fondamentale en cybernétique; sa signification exacte sera étudiée en détail plus tard. Tout système existant en réalité se compose d'objets concrets: dispositifs techniques, personnes commandant ces derniers, ressources matérielles, etc. Ces objets sont liés entre eux et à l'environnement par des liaisons déterminées: forces, courants d'énergie, de substances, d'information. Or, la cybernétique fait abstraction du contenu physique des propriétés des objets et des liaisons; elle considère le système réel comme un ensemble abstrait d'éléments possédant des propriétés communes et liés les uns aux autres par des relations déterminées par la nature des liaisons existantes. Une pareille représentation permet d'abandonner l'habituelle division des systèmes en systèmes mécaniques, électriques, chimiques, biologiques, etc., et d'introduire la notion de *système cybernétique abstrait* désignant l'ensemble d'éléments liés entre eux et se trouvant en interaction. On voit sur la figure I-1 un exemple de système cybernétique à quatre éléments et à six liaisons réciproques.

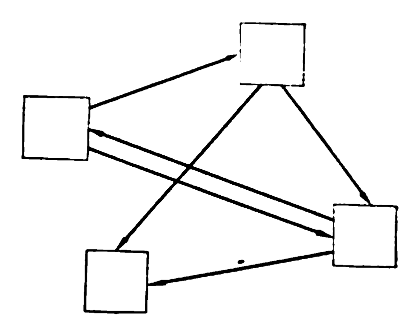


Fig. I-1. Exemple d'un système cybernétique

La représentation du système sous la forme d'un ensemble d'éléments permet de donner sa définition mathématique en termes de la théorie des ensembles. Dans certains cas fort importants, les liaisons existant entre les éléments se laissent définir aisément par les moyens de la logique mathématique. Aussi ces deux disciplines — la théorie des ensembles et la logique mathématique — se trouvent-elles à la base de la définition mathématique du système; elles constituent pour cette raison les chapitres liminaires de cet ouvrage.

Les systèmes que l'on rencontre dans la pratique se divisent, en fonction de leur structure et du caractère des liaisons, en systèmes déterministes et systèmes probabilistes. On dit que le système est *déterministe* quand on connaît avec précision les lois de son mouvement et son comportement à l'avenir. En ce qui concerne un système *probabiliste*, son comportement ne peut jamais être prédit exactement a priori. Un mouvement d'horlogerie est un système déterministe; par contre, les systèmes de contrôle statistique de la production, les systèmes d'arrivée des navires aux ports maritimes, le stock des marchandises dans un magasin desservant un grand nombre de fournisseurs et d'acheteurs sont des systèmes probabilistes.

Les problèmes traités par la cybernétique exigent dans la plupart des cas l'étude d'assez complexes systèmes probabilistes, composés d'un grand nombre d'éléments et présentant des liaisons internes variées et ramifiées. Tels sont notamment la plupart des systèmes de production, des systèmes économiques, sociaux, biologiques. La définition mathématique de ces systèmes nécessite l'emploi très large des méthodes relevant de la théorie des probabilités et de la statistique mathématique, sans parler de la théorie des ensembles et de la logique mathématique.

Nous n'avons mentionné jusque-là que les méthodes mathématiques employées pour définir les systèmes cybernétiques. Or, le but de la cybernétique est la commande des systèmes, ou la gestion. Pour rechercher les moyens de résolution de ce problème, il est nécessaire de se faire une idée bien nette de ce qu'est la « gestion ».

Au sens large, on entend par *gestion* l'activité menée pour gouverner le travail de l'autrui orienté vers des buts précis. Le processus de la gestion consiste à prendre des décisions sur les actions les plus raisonnables dans le contexte de la situation existante. Celui qui effectue la gestion prend les décisions après avoir apprécié la situation sur la base de l'information reçue par ses sens, ou à l'aide d'instruments de mesure, ou par l'entremise d'autres personnes. Bien souvent, l'information reçue s'avère insuffisante pour apprécier la situation de façon univoque. Alors l'homme fait appel à son expérience, à ses connaissances, à sa mémoire, à son intuition. Une qualité remarquable de l'homme est justement sa faculté de prendre les décisions dans des circonstances très incertaines.

Or, dans les grandes entreprises modernes, même un administrateur très expérimenté ne peut réaliser une gestion efficace s'il s'appuie exclusivement sur ses connaissances et sur son intuition. Cet état de choses est la cause des carences fréquentes dans le fonctionnement des entreprises importantes: le travail par saccades; l'impossibilité d'assurer le ravitaillement normal en matières premières et matériaux sans gonfler démesurément les stocks; les problèmes de transport qui s'aggravent, etc.

La cybernétique se donne pour tâche de faciliter à l'homme la prise des décisions importantes, en confiant à des machines automatiques la collecte et le traitement d'énormes quantités d'information relative à l'état de la production, l'analyse des situations existantes et l'élaboration des recommandations sur les actions les plus avantageuses. Les dispositifs automatiques assumant l'ensemble de ces opérations sont désignés sous le terme de *systèmes de gestion automatiques*. Ces systèmes se basent sur l'emploi des machines à calcul numériques, ou ordinateurs.

Le rôle des ordinateurs en cybernétique est tellement important qu'il convient ,pen parler d'une façon plus détaillée.

Au début, les ordinateurs ne s'employaient que pour faire des calculs traditionnels en quelques secondes au lieu de nombreuses heures. Or, on s'est vite rendu compte que l'accélération vertigineuse des opérations de calcul recèle des traits nouveaux. Auparavant, de l'infinité des solutions possibles, le projeteur ou l'économiste ne pouvait analyser que quelques-unes, celles qu'il a jugées (pour une raison quelconque) les plus dignes d'attention; aujourd'hui au contraire, il se trouve en mesure de les étudier toutes et d'en choisir la meilleure. C'est là que prennent naissance les idées d'optimisation, qui donneront naissance plus tard à toute une série de branches nouvelles des mathématiques.

On a découvert ensuite que l'ordinateur utilisé dans une entreprise de production assume sans peine le traitement des masses importantes d'information et représente donc un collaborateur précieux de l'homme dans le processus de la gestion.

Pour pouvoir utiliser les ordinateurs dans les buts de la gestion, on a eu besoin des méthodes mathématiques permettant d'analyser l'information parvenue, d'éliminer les informations inutiles et de mettre en valeur l'information la plus essentielle, d'employer cette information pour apprécier la situation existante et d'élaborer les recommandations propres à assurer la réalisation optimale des buts de gestion fixés. La nécessité de résoudre les problèmes de ce genre a fait naître de nouvelles branches des mathématiques, telles que théories de l'information, des jeux, des décisions statistiques, des files d'attente, programmation linéaire et dynamique, etc. Quelques-unes de ces méthodes mathématiques modernes seront exposées dans ce livre; d'autres font l'objet des cours appropriés.

I-2. TRANSMISSION ET CODAGE DE L'INFORMATION

Les liaisons existant entre les éléments d'un système peuvent servir à des fins différentes, par exemple à transmettre l'énergie, les substances, les efforts, etc. Or, dans les systèmes cybernétiques, c'est le contenu informationnel des liaisons qui nous intéresse en premier lieu, ou, en d'autres mots, la possibilité d'utiliser les liaisons pour transmettre les données sur les différents états des éléments du système.

Toute information concernant un événement quelconque se produisant à l'intérieur ou à l'extérieur du système est désignée en technique sous le terme de *message*. Les liaisons informationnelles servant à transmettre des messages sont appelées *canaux de communication*. Les supports physiques de l'information circulant dans les canaux de communication sont les *signaux*. On voit sur la figure I-2 le schéma fonctionnel d'un canal de communication.

Le mot « signal », provenant du lat. *signum*, désigne les signes conventionnels électriques, sonores, optiques et autres destinés

à transmettre les messages. Généralement, la nature des signaux fait l'objet d'une convention entre la personne désirant transmettre l'information et la personne qui la reçoit; elle n'a pas de liaison directe avec le contenu de l'information véhiculée. Aussi peut-on aisément transformer les signaux d'une nature en ceux de nature différente sans changer le contenu de l'information et conserver ces signaux (après les avoir mis, par exemple, sous la forme d'une notation littérale) afin d'utiliser à l'avenir l'information qu'ils portent.

Considérons, à titre d'exemple, les signaux utilisés en télégraphie. Le message initial écrit sur l'imprimé postal est transcrit à l'aide

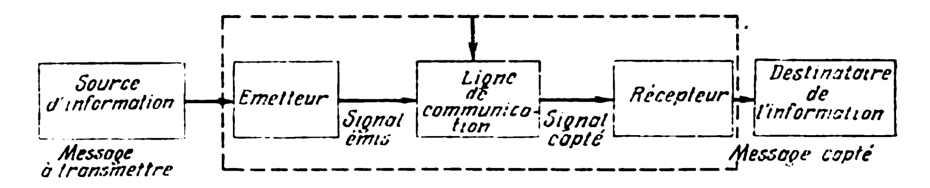


Fig. I-2. Schéma fonctionnel d'un canal de communication

du code Morse formé de points et de traits et est envoyé ensuite par la ligne de communication sous la forme d'impulsions de courant de courte et de grande durée. D'après ces impulsions, le récepteur reconstitue le contenu du texte. Nous voyons donc que le message existe dans le cas considéré sous la forme de signaux de forme différente: un texte littéral, points et traits du code Morse, impulsions s de courant dans la ligne de communication, etc. L'action de faire passer le signal d'une forme à une autre s'appelle *codage*.

La communication se présente donc comme suit. On crée tout d'abord un ensemble de symboles: lettres, mots, points, traits, etc., dont le sens est connu tant de l'auteur du message que du récepteur. Cet ensemble de symboles porte le nom d'*alphabet*. Les symboles eux-mêmes sont précisés par convention.

La thèse fondamentale de la théorie des communications stipule que les symboles constituant un alphabet ne peuvent être variés à l'infini. Pour cette raison la transmission des messages de toute sorte se réalise avec un nombre restreint de symboles. C'est ainsi que tous les messages possibles en langue russe se construisent avec un alphabet de 33 lettres.

Au cours de la communication, le transmetteur choisit dans son alphabet les symboles un à un, les transforme en signaux correspondants et les envoie par le canal de communication. Dans ce dernier, les signaux subissent l'action de parasites et se déforment, ce qui fait que les signaux captés par le récepteur diffèrent des signaux envoyés.

Le processus de la réception consiste en ce que le destinataire, après avoir reçu un signal quelconque, l'identifie à l'un des symboles de l'alphabet, c.-à-d. exclut tous les symboles sauf un. Ce problème se complique considérablement lorsque les signaux subissent de fortes déformations pendant leur passage par le canal de communication. Ce sont les méthodes de lutte contre ces difficultés qui constituent l'essentiel de la théorie des communications.

Les alphabets utilisés dans les systèmes de communication techniques sont assez variés. Cependant, pour certaines raisons, on adopte très souvent l'*alphabet binaire* n'employant que deux espèces de symboles, représentés conventionnellement par les chiffres 0 et 1, de sorte que tout message se présente comme une succession de zéros et d'unités, par exemple 100110100.

Il est facile de calculer le nombre total de messages composés de m lettres d'alphabet binaire; ce nombre est égal à 2^m. En particulier, n'importe quelle lettre de l'alphabet français peut être représentée au moyen de six signes de l'alphabet binaire, par exemple, $a = 000001$, $b = 000010$, $c = 000011$, etc. Puisque six signes binaires permettent $2^6 = 64$ combinaisons différentes de symboles, on a la possibilité d'exprimer non seulement toutes les lettres de l'alphabet mais aussi les signes de ponctuation. Par conséquent, l'alphabet binaire permet de représenter et de transmettre par le canal de communication n'importe quel message littéral.

L'alphabet binaire convient aussi pour la transmission des données numériques; cependant, pour le faire, on est amené à utiliser des systèmes de numération particuliers.

Dans le système de numération décimale universellement répandu les différents nombres s'écrivent au moyen de dix chiffres 0, 1, . . ., 9 disposés dans un ordre déterminé et prenant des valeurs différentes en fonction du rang occupé par chaque chiffre. Par exemple, l'écriture 395 correspond au nombre déterminé par la relation

$$3 \cdot 10^2 + 9 \cdot 10^1 + 5 \cdot 10^0.$$

Le nombre 10 est dit *base* de la numération.

De façon analogue, on écrira n'importe quel nombre N à l'aide d'un autre système de numération. ayant pour base n'importe quel nombre entier R, au moyen de chiffres dont la quantité est égale à la base de la numération. L'écriture $\ldots d_3d_2d_1d_0$, où d_i sont les chiffres du nombre N $(0 \leqslant d_i < R)$ détermine alors la quantité

$$N = \ldots d_3R^3 + d_2R^2 + d_1R^1 + d_0R^0. \qquad \text{(I-1)}$$

Ainsi, en système de numération octal (employé dans certaines machines à calcul numériques), le nombre 395 peut être représenté comme suit:

$$6 \cdot 8^2 + 1 \cdot 8^1 + 3 \cdot 8^0,$$

donc s'écrira comme 613.

Si l'on adopte l'alphabet binaire, un nombre ne s'écrira qu'avec les symboles 0 et 1. Cette écriture est possible avec le système de numération binaire de base 2. N'importe quel nombre de 0 à 15 peut s'écrire dans le système binaire à l'aide d'un nombre à quatre chiffres :

$$3 = 0\cdot 2^3 + 0\cdot 2^2 + 1\cdot 2^1 + 1\cdot 2^0, \text{ c.-à-d. } 0011;$$
$$5 = 0\cdot 2^3 + 1\cdot 2^2 + 0\cdot 2^1 + 1\cdot 2^0, \text{ c.-à-d. } 0101;$$
$$9 = 1\cdot 2^3 + 0\cdot 2^2 + 0\cdot 2^1 + 1\cdot 2^0, \text{ c.-à-d. } 1001.$$

Naturellement, on peut omettre d'écrire les zéros dans les rangs supérieurs de ces nombres, si bien que les nombres 3 et 5 s'écriront comme 11 et 101. Le nombre 395, transcrit avec la base 2:

$$395 = 2^8 + 2^7 + 2^3 + 2^1 + 2^0,$$

s'écrira dans le système binaire au moyen d'un nombre à neuf chiffres 110001011.

L'inconvénient du système binaire est que l'écriture des grands nombres exige une quantité exagérée de chiffres, ce qui complique la lecture des nombres et l'appréciation rapide de leur ordre de grandeur. Aussi emploie-t-on souvent des systèmes de numération de base mixte, par exemple le système décimal-binaire dans lequel le nombre lui-même s'écrit en décimal et les chiffres de ses différents rangs, en binaire, avec utilisation de quatre rangs binaires pour chaque chiffre décimal. Le nombre 395 se présente en décimal-binaire comme suit :

$$\underbrace{0011}_{3}\ \underbrace{1001}_{9}\ \underbrace{0101}_{5}$$

I-3. NOTION DE SYSTÈME COMMANDÉ

Conformément à la définition de la gestion (ou de la commande) donnée ci-dessus, on peut représenter tout système commandé comme l'ensemble de deux parties : une partie commandée, dite *objet* de la commande, et une partie qui effectue la commande, dite *organe de commande* ou *opérateur* [6].

L'objet de commande se caractérise par :

1) un but, ou la possibilité de fournir un résultat utile ;

2) un état, exprimé par des mouvements concrets et susceptible de varier par suite de variations des conditions extérieures dans lesquelles l'objet est placé ;

3) la commandabilité, c.-à-d. la faculté de réagir aux sollicitations extérieures appliquées à ses organes appropriés, dits organes de commande.

La mission de l'opérateur consiste à agir sur les organes de commande de façon que l'objet réalise son but. A cet effet, on doit établir

entre l'opérateur et l'organe de commande un échange d'information réciproque. L'information en question est portée, d'une part, par les signaux de commande à l'aide desquels l'opérateur agit sur l'objet de commande: c'est l'*information de commande*, et d'autre part, par les données sur l'état de l'objet commandé à la base desquelles l'opérateur choisit les signaux de commande: c'est l'*information en retour*.

Le schéma fonctionnel de la figure I-3 montre les courants d'information essentiels dans le système commandé. Outre les courants

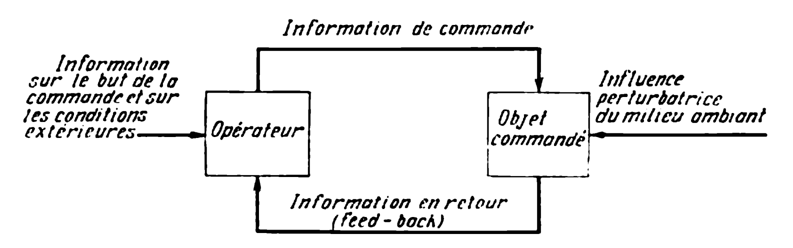


Fig. I-3. Circulation de l'information dans un système commandé

de circulation de l'information de commande et de l'information en retour, on y distingue la liaison de l'objet commandé et de l'opérateur avec le milieu ambiant. Toute variation des conditions extérieures exerce une influence directe sur l'objet commandé en modifiant le caractère de son mouvement et en l'empêchant d'atteindre son but. Pour que la commande de l'objet soit efficace, il faut que l'opérateur dispose de l'information sur les conditions extérieures et puisse en tenir compte à l'élaboration des signaux de commande; il doit encore connaître le but poursuivi par la commande.

L'information de commande résulte du traitement de toutes les catégories de l'information arrivant à l'opérateur. Une partie d'information peut être conservée pour être utilisée dans l'avenir. Les fonctions d'opérateur peuvent être remplies tant par l'homme que par un dispositif artificiel, soit mécanique, soit électronique. Aujourd'hui le rôle d'opérateur est souvent confié à des ordinateurs, universels ou spécialisés.

Dans l'objet commandé, il se produit le traitement de l'information de commande, qui se manifeste par la variation du caractère du mouvement de l'objet. Sous la forme de l'information en retour, ces variations sont transmises à l'opérateur par le canal de liaison en retour.

La *liaison en retour* (ou *réaction*, ou *feed-back*) joue un rôle fondamental dans la commande efficace, car elle permet à l'opérateur d'apprécier à tout moment le degré de réalisation du but fixé et d'élaborer en conséquence les signaux de commande de la façon la

plus rationnelle. Aussi le principe de la réaction est-il mis à la base de la plupart des processus de commande et, en particulier, à la base de presque toutes les activités de l'homme. Par exemple, quand il tend sa main pour prendre un crayon sur la table, l'homme effectue une comparaison permanente, quoique inconsciente, des positions mutuelles de la main et du crayon, grâce à quoi le mouvement prend un caractère remarquablement net.

Le principe de la réaction est étudié en détail dans le cours de « Théorie de la commande automatique ». Le présent cours, lui, vise à étudier les méthodes de définition mathématique des systèmes de commande automatique et de mettre en valeur les procédés et les techniques de la prise de décision mis à la base des dispositifs de commande.

A l'aide de quelques exemples de systèmes de nature physique différente, essayons de rendre plus clair le schéma général du processus de la commande.

Exemple I-1. Régulateur centrifuge des tours d'une machine à vapeur (fig. I-4). L'opérateur du système considéré est le régulateur centrifuge des tours. L'organe de commande est la vanne réglant la pénétration de la vapeur dans la

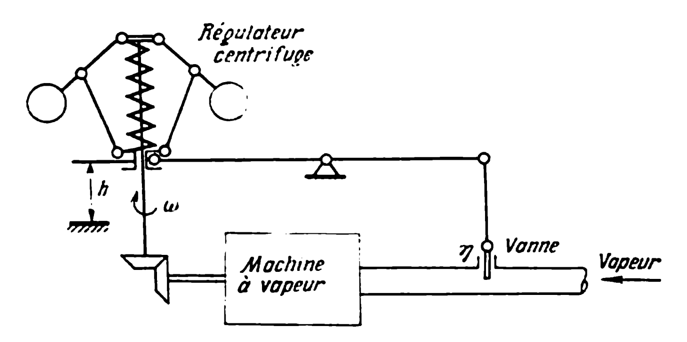


Fig. I-4. Schéma de régulation de la vitesse d'une machine à vapeur

machine. Avant le commencement du travail, on doit introduire dans le régulateur l'information sur le but de la commande, ce qu'on fait en affichant sur le régulateur le nombre des tours à maintenir ω_0 par la disposition appropriée des masselottes et par la variation de la tension des ressorts. Le régulateur centrifuge transforme l'information en retour sur le nombre de tours réel ω de la machine en variation de la hauteur h du manchon ; cette information de commande, par l'intermédiaire du levier, se transforme en information sur la position de la vanne qui modifie l'arrivée de la vapeur et fait varier, par là même, les tours de la machine. L'influence du milieu ambiant se manifeste ici par la variation de la charge sur la machine à vapeur, ce qui fait changer la vitesse de rotation.

Exemple I-2. Dispatcher de chemin de fer. Le rôle d'opérateur est confié ici à un homme, le dispatcher, qui, se trouvant dans le dispatching, reçoit par intercom ou par téléphone l'information en retour sur l'état des trains dans le secteur qui lui est confié. Cette information se présente sous la forme de messages

concernant la préparation d'un train à la mise en marche, le retard d'un train par rapport à l'horaire établi, la nécessité de former un train spécial (non prévu par l'horaire), l'occupation de la voie, etc. La mission du dispatcher est de dépouiller l'information reçue et de prendre les décisions propres à assurer l'observation la plus stricte de l'horaire des trains. Cette information de commande s'exprime en ordres adressés aux conducteurs de trains ou aux chefs de gares et prescrivant de changer le trafic (accélérer, retarder, garer un train, etc.).

Exemple I-3. Guidage d'un missile antiaérien. En qualité d'exemple d'un système automatique desservi par une machine à calculer électronique, on donne

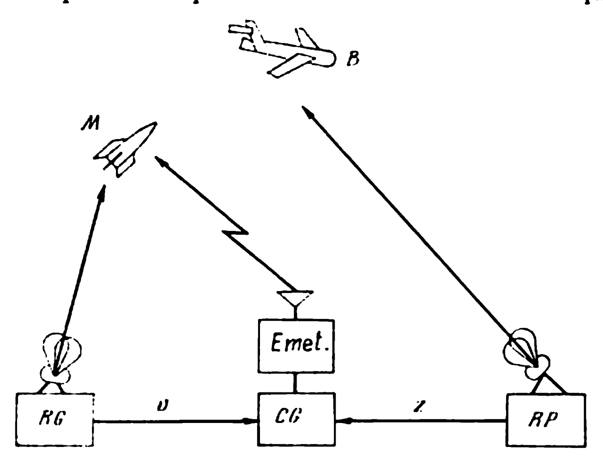


Fig. I-5. Système de guidage d'un missile antiaérien

sur la figure I-5 le schéma fonctionnel de guidage d'un missile antiaérien sur un but mobile. Dans ce système, l'information en retour est contenue en signaux délivrés par le radar de poursuite RP déterminant les éléments de position du but et par le radar de guidage RG déterminant les éléments de position du missile lancé pour intercepter le but. Le calculateur de guidage CG, à la base de cette information, élabore les signaux qui, émis par radio et captés ensuite par l'organe de direction du missile, corrigent le vol de celui-ci en fonction de la position réciproque du missile et du but.

PROBLÈMES À L'INTRODUCTION

I-1. Le système comporte sept éléments dont chacun est lié à tous les autres par des liaisons élémentaires bistables, c.-à-d. n'ayant que deux états: « la liaison existe » et « la liaison n'existe pas ». Quel est le nombre des états possibles d'un pareil système? En combien de temps pourra-t-on analyser tous les états possibles possédant la propriété P si l'étude expérimentale d'un seul état dure 1 seconde?

I-2. Supposons que le système considéré en I-1 ne comporte que des liaisons dirigées successives (fig. I-6, *a*) ou qu'il a la structure hiérarchique (fig. I-6,*b*). De combien se réduira le nombre de tous les états possibles de ce système? Citer quelques exemples de systèmes ayant les structures des deux types.

I-3. Le polynôme (I-1) servant à représenter un nombre en système de numération de base R peut être écrit sous la forme

$$N = \{[(\ldots d_3) R + d_2] R + d_1\} R + d_0, \qquad \text{(I-2)}$$

ce qui équivaut à la succession de formules

$$N = N_1R + d_0, \quad N_1 = N_2R + d_1, \quad N_2 = N_3R + d_2 \ldots$$

Utilisant ces formules, argumenter la règle suivante de passage du nombre 395 du système décimal au système octal:

395 : 8 = 49 reste 3 ou en abrégé	395	
49 : 8 = 6 reste 1	49	3
6 : 8 = 0 reste 6	6	1
	0	6,

ce qui donne le nombre octal 613.

Comment peut-on définir la règle de passage d'un nombre du système décimal au système binaire?

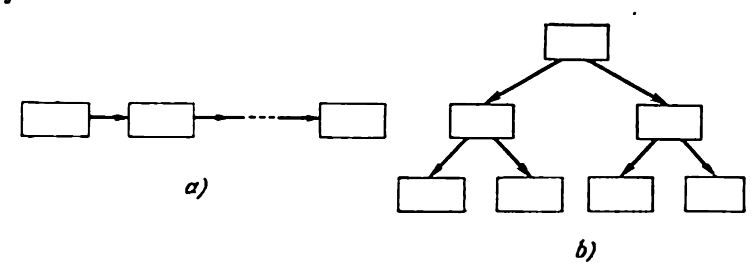


Fig. I-6. Systèmes aux liaisons successives (*a*) et à structure hiérarchique (*b*)

I-4. Ecrire en systèmes binaire, octal et sexdécimal les nombres décimaux suivants (qui sont donnés en système de numération décimal):

27, 467, 519, 1263.

N o t a. Pour les chiffres de 10 à 15 du système sexdécimal, employer les lettres de l'alphabet latin *a*, *b*, *c*, *d*, *e*, *f*.

I-5. Ecrire les nombres du problème I-4 en décimal-binaire et en octal-binaire. Combien de rangs binaires doit-on occuper pour écrire un chiffre octal? Lequel de ces deux systèmes est plus économique au point de vue de la quantité de nombres binaires utilisés?

I-6. Etablir s'il existe le circuit de réaction dans les systèmes de régulation de la circulation routière:

1) par les feux rouge, orange et vert s'allumant à tour de rôle pour une durée établie à l'avance;

2) par un agent de circulation.

I-7. Trouver le circuit de réaction dans les schémas des figures I-4 et I-5.

INDEX DES NOTATIONS

$\{\ldots\}$ — ensemble
$\in$ — appartenance d'un élément à un ensemble
$\notin$ — non-appartenance à un ensemble
$\subseteq$ — symbole d'inclusion
$\subset$ — symbole d'inclusion stricte
$\cup$ — réunion d'ensembles
$\cap$ — intersection d'ensembles
$\setminus$ — différence d'ensembles
$\times$ — produit direct d'ensembles
$\overline{X}$ — complémentaire de l'ensemble X
X^s — puissance de l'ensemble
$\varnothing$ — ensemble vide
I — ensemble universel
R — ensemble des nombres réels
$\mathfrak{M}$ — système d'ensembles
$\sup M$ — borne supérieure de l'ensemble M
$\inf M$ — borne inférieure de l'ensemble M
$\mathrm{pr}_i\, M$ — projection de l'ensemble sur l'axe des i
$(a_1, \ldots, a_n)$ — ensemble ordonné (cortège, vecteur)
Λ — cortège vide
$\forall$ — quantificateur universel
$\rightleftarrows$ — équivalence logique
$\rightarrow$ — inférence, application
$f \circ g$ — composée des fonctions f et g
$\equiv$ — symbole de relation d'équivalence
$\leqslant$ — symbole de relation d'ordre
$<$ — symbole de relation d'ordre stricte
$\gg$ — symbole de relation de dominance
$d(x, y)$ — distance des éléments d'un ensemble
$\|x\|$ — norme de x
E_n — espace euclidien à n dimensions
$C_{[a, b]}$ — espace des fonctions continues
x/y — fonction de Sheffer
$\bar{x}$ — inversion de la proposition x
$\oplus$ — addition modulo deux
Z — espace des épreuves
z — épreuve liée à l'expérience, variable aléatoire
$\overset{\circ}{z}$ — variable aléatoire centrée
$p(z),\ z \in Z$ — distribution des probabilités sur l'espace Z
$P(S),\ P_S$ — probabilité de l'événement
$p(z/S)$ — distribution conditionnelle des probabilités
$p(y, z)$ — distribution commune des probabilités
$\bar{z},\ \nu,\ M(z),\ \overline{f(z)}$ — valeur moyenne
$\sigma,\ \sigma_z$ — écart quadratique moyen

$R(\nu, \omega)$ — distribution uniforme
$N(\nu, \sigma^2)$ — distribution normale
$w(r, n, p)$ — distribution binomiale
$w(r, a)$ — distribution de Poisson
u — commande
ϑ — état de la nature
x — état du système
$T(x, u)$ — transformation de l'état du système
$q(x, u)$, $Q(x, u)$ — fonction objectif
$J(u)$, $J_n(u)$ — critère de qualité de la commande
ξ, η — stratégies mixtes
$L(\xi, \eta)$ — fonction de pertes
A — espace des décisions
a — décision, action
a^* — action de Bayes
$R^*(\xi)$ — pertes correspondant à l'action de Bayes
$d(z)$ — fonction de décision
$\rho(\vartheta, d)$ — fonction de risque
$\rho^*(\xi)$ — risque de Bayes

PREMIÈRE PARTIE

ÉLÉMENTS DE MATHÉMATIQUES DISRÈTES

CHAPITRE PREMIER

NOTIONS ESSENTIELLES

DE LA THÉORIE DES ENSEMBLES

1-1. ENSEMBLES FINIS ET INFINIS

a) Définitions principales

La notion d'ensemble [7 à 9] est une des entités mathématiques difficilement assignables à l'aide de notions élémentaires, aussi devra-t-on se borner à lui donner une explication descriptive. On entend par *ensemble* une collection d'objets bien distincts considérés comme formant un tout.

On peut parler de l'ensemble des chaises garnissant une pièce; de l'ensemble des habitants de la ville de Riazan ; de l'ensemble des étudiants d'un groupe; de l'ensemble des entiers naturels; de l'ensemble des lettres d'un alphabet; de l'ensemble des états d'un système, etc. Dans tous ces cas il ne s'agira d'un ensemble que lorsque les éléments de cet ensemble sont différenciables. Par exemple, on ne peut parler de l'ensemble des gouttes d'eau contenues dans un verre, vu qu'il n'est pas possible de définir nettement chaque goutte isolée.

Les objets isolés constituant l'ensemble sont dits *éléments* de l'ensemble. C'est ainsi que le nombre 3 est un élément de l'ensemble des entiers naturels; la lettre b est un élément de l'ensemble des lettres de l'alphabet latin.

Pour désigner d'une manière générale un ensemble, on emploie les accolades $\{\ \}$ à l'intérieur desquelles on énumère les éléments de l'ensemble. Les ensembles concrets sont désignés par des lettres capitales A, S, X, . . . ou par des lettres majuscules munies d'indices A_1, A_2, . . . Pour désigner sous une forme générale les éléments d'un ensemble, on utilise des lettres minuscules a, s, x, . . . ou des lettres minuscules affectées d'indices a_1, a_2, . . .

Pour dire que l'objet a est élément de l'ensemble S, on écrira $a \in S$; cette écriture se prononce « a est élément de S » ou « a appartient à S ». Pour dire, au contraire, que a n'est pas élément de l'ensemble S, on écrira $a \notin S$. La notation $x_1, x_2, \ldots, x_n \in S$ sera employée pour abréger l'écriture $x_1 \in S$, $x_2 \in S$, . . ., $x_n \in S$.

Il y a des ensembles finis et infinis. On dit que l'ensemble est *fini* si le nombre de ses éléments est fini, c.-à-d. s'il existe un nombre naturel N qui est le nombre des éléments de l'ensemble. Un ensemble est dit *infini* quand il contient un nombre infini d'éléments.

Pour pouvoir opérer avec des ensembles concrets, on doit savoir les définir. Il existe deux modes de définition d'un ensemble : énumération et description. La définition énumérative d'un ensemble consiste à énumérer tous les éléments qui le forment. C'est ainsi que pour définir l'ensemble des étudiants d'un groupe qui ont toutes les notes excellentes, on peut nommer tous les étudiants qui satisfont à cette condition, par exemple : {Ivanov, Pétrov, Sidorov}. Ce mode de définition convient lorsqu'on a affaire à des ensembles finis de peu d'éléments ; or, il peut servir quelquefois à définir aussi bien des ensembles infinis, tels que $\{2, 4, 6, 8, \ldots\}$. On conçoit qu'une notation pareille n'est valable que si l'on comprend clairement la signification des points de suspension.

La définition descriptive d'un ensemble revient à indiquer une propriété caractéristique commune à tous les éléments de l'ensemble. Par exemple, si M est l'ensemble des étudiants du groupe, l'ensemble A des étudiants de ce groupe qui ont toutes les notes excellentes s'écrira comme suit :

$$A = \{x \in M : x \text{ est un étudiant dont toutes les notes sont excellentes}\},$$

expression que l'on prononce : ensemble A des éléments x définis sur l'ensemble M par la condition d'après laquelle x est un étudiant du groupe qui n'a que des notes excellentes.

Dans les cas où il est très clair sur quel ensemble sont définis les éléments x, on omet d'indiquer que x appartient à M et l'on écris

$$A = \{x : x \text{ est un étudiant dont toutes les notes sont excellentes}\}.$$

Citons quelques exemples de définition descriptive de différents ensembles :

$\{x : x \text{ est pair}\}$ est l'ensemble des nombres pairs ;

$\{x : x^2 - 1 = 0\}$ est l'ensemble $\{+1, -1\}$.

Soit C l'ensemble des nombres entiers. Alors $\{x \in C : 0 < x \leqslant 7\}$ est l'ensemble $\{1, 2, 3, 4, 5, 6, 7\}$.

Un concept très important de la théorie des ensembles est celui d'ensemble vide. On dit que l'ensemble est *vide* quand il ne contient pas un seul élément. L'ensemble vide se désigne par le symbole $\varnothing$. Par exemple,

$$\{x \in C : x^2 - x + 1 = 0\} = \varnothing.$$

La notion d'ensemble vide joue un rôle très important lorsqu'il s'agit de définir un ensemble par description. Sans cette notion,

on ne pourrait parler de l'ensemble des étudiants du groupe qui ont toutes les notes excellentes, ou de l'ensemble des racines réelles d'une équation du second degré, sans voir au préalable s'il y a effectivement, dans le groupe donné, au moins un étudiant dont toutes les notes sont excellentes, ou que l'équation proposée admet des racines réelles. Faisant intervenir la notion d'ensemble vide, on peur opérer en toute tranquillité avec l'ensemble des étudiants du groupe qui ont toutes les notes excellentes, sans se soucier nullement du fait que ce groupe présente ou non des étudiants aussi brillants. Par convention, l'ensemble vide se range parmi les ensembles finis.

Passons maintenant au problème d'égalité des ensembles. On dit que deux ensembles sont *égaux* quand ils sont formés des mêmes éléments, c.-à-d. quand ils représentent un seul et même ensemble. Deux ensembles X et Y ne sont pas égaux ($X \neq Y$) quand il y a dans X des éléments qui n'appartiennent pas à Y, ou quand il y a dans Y des éléments qui n'appartiennent pas à X. On voit sans peine que pour des ensembles quelconques X, Y et Z on a:

$$X = X;$$

$$\text{si } X = Y, \text{ alors } Y = X;$$

$$\text{si } X = Y \text{ et } Y = Z, \text{ alors } X = Z.$$

De la définition de l'égalité des ensembles il découle que l'ordre des éléments dans l'ensemble n'a pas d'importance. C'est ainsi que les ensembles $\{3, 4, 5, 6\}$ et $\{4, 5, 6, 3\}$ représentent un seul et même ensemble.

Considérant des ensembles de différente nature, on est souvent amené à parler du nombre des éléments d'un ensemble. Pour que cette notion ait un sens bien déterminé, il faut préciser qu'un ensemble ne peut contenir d'éléments identiques. Une notation $\{2, 2, 3, 5\}$ sera considérée comme incorrecte et sera remplacée par la notation $\{2, 3, 5\}$. C'est ainsi que l'ensemble des diviseurs premiers du nombre 60 est $\{2, 3, 5\}$.

b) Notion de sous-ensemble

Un ensemble X est sous-ensemble de l'ensemble Y si tout élément de X appartient aussi à Y. Soient Y l'ensemble des étudiants d'un groupe et X l'ensemble des étudiants du même groupe qui ont toutes les notes excellentes. Du fait que l'étudiant dont toutes les notes sont excellentes est en même temps étudiant du groupe donné, l'ensemble X est sous-ensemble de l'ensemble Y.

On formule aisément de nombreuses définitions relevant de la théorie des ensembles sous la forme d'expressions mathématiques contenant quelques symboles logiques. La définition du sous-ensemble sera donnée en employant deux symboles suivants:

$\forall$ symbole appelé quantificateur et signifiant « n'importe quel » « quel que soit », « pour tout »;

$\rightarrow$ symbole de l'inférence signifiant « entraîne ».

Un sous-ensemble se laissant formuler de la sorte: pour tout x, l'assertion « x appartient à X » entraîne l'assertion « x appartient à Y », s'écrira comme suit:

$$\forall x : x \in X \rightarrow x \in Y. \quad (1\text{-}1)$$

L'assertion « X est sous-ensemble de Y » s'écrit brièvement

$$X \subseteq Y \quad (1\text{-}2)$$

et se prononce « Y inclut X », ou « X est contenu dans Y ». Le symbole $\subseteq$ ci-dessus signifie l'inclusion. Pour souligner que Y contient aussi des éléments autres que ceux de X, on fait appel au symbole d'inclusion stricte $\subset$:

$$X \subset Y. \quad (1\text{-}3)$$

La relation entre les symboles $\subset$ et $\subseteq$ se définit par

$$X \subset Y \rightleftarrows X \subseteq Y \text{ et } X \neq Y. \quad (1\text{-}4)$$

Le symbole $\rightleftarrows$ employé ici indique l'équivalence (au sens de « la même chose que »).

Notons quelques propriétés du sous-ensemble qui résultent de sa définition:

$$X \subseteq X \text{ (réflexivité)};$$

$$[X \subseteq Y \text{ et } Y \subseteq Z] \rightarrow X \subseteq Z \text{ (transitivité)}.$$

Un peu moins évidente est la propriété selon laquelle, pour tout ensemble M, on a

$$\varnothing \subseteq M. \quad (1\text{-}5)$$

En effet, l'ensemble vide $\varnothing$ ne contient aucun élément. Donc, ajoutant à M un ensemble vide, on n'y ajoute en réalité rien. Aussi peut-on dire toujours qu'un ensemble quelconque M contient un ensemble vide à titre de sous-ensemble.

c) Borne supérieure et borne inférieure d'un ensemble

Si l'ensemble avec lequel on opère est celui des nombres réels, on peut comparer les éléments de cet ensemble au point de vue de leur grandeur. Assez souvent, il devient alors nécessaire de voir quel élément de l'ensemble est le plus grand ou le plus petit. Pour un ensemble fini le problème n'est pas difficile. Ainsi, pour l'ensemble $T = \{3, 4, 5, 6\}$ $\min T = 3$, $\max T = 6$. Il en est quelquefois autrement pour des ensembles infinis.

Soient R l'ensemble de tous les nombres réels, et $S = \{x \in R: m < x < M\}$. L'ensemble S, représentant un segment non fermé de l'axe réel, n'a ni plus grand ni plus petit élément. Cependant, il est possible qu'un ensemble de ce type ait des bornes: ce seront dans notre cas les nombres m et M qui, étant ajoutés à l'ensemble S, le complètent jusqu'à un segment fermé. Le point M porte le nom de supremum de S, $M = \sup S$, et le point m porte le nom d'infimum de S, $m = \inf S$.

Théorème 1-1 (théorème de la borne supérieure et de la borne inférieure d'un sous-ensemble). *Si* $B \subseteq A$, *alors*

$$\inf B \geqslant \inf A; \quad \sup B \leqslant \sup A. \tag{1-6}$$

D é m o n s t r a t i o n. Soit b' le plus petit élément de B, de sorte que $b' \in B$ et $b' = \inf B$. Or, $B \subseteq A$, ce qui fait que $b' \in A$. Soit a' un élément de A tel qu'il ait la valeur la moins élevée, en sorte que $a' \in A$ et $a' = \inf A$. Dans ce cas, si $b' = a'$, on a $b' = \inf A$; si $b' \neq a'$, on a $b' > a' = \inf A$. Ainsi donc, $b' \geqslant \inf A$ ou $\inf B \geqslant \inf A$.

On démontre de façon analogue la seconde assertion du théorème.

1-2. OPÉRATIONS SUR LES ENSEMBLES

a) Considérations préliminaires

Les ensembles se prêtent à des opérations qui rappellent beaucoup les opérations d'addition et de multiplication en algèbre élémentaire. Pour mieux comprendre les opérations effectuées sur les ensembles, il est nécessaire de se rappeler les lois régissant l'algèbre élémentaire.

Soient a et b deux nombres, $a + b$ leur somme et ab leur produit. La somme et le produit possèdent certaines propriétés, appelées aussi lois d'algèbre:

1. $a + b = b + a$; $ab = ba$. C'est la loi commutative.
2. $(a + b) + c = a + (b + c)$; $(ab)\,c = a\,(bc)$. C'est la loi associative.
3. $(a + b)\,c = ac + bc$. C'est la loi distributive.

Notons qu'en ce qui concerne les lois associative et commutative, il est possible de substituer la multiplication à l'addition, et vice versa. On aura obtenu une autre loi, tout aussi valable que la première. Par contre, la loi distributive ne permet pas une symétrie pareille. Substituant dans cette loi la multiplication à l'addition et l'addition à la multiplication, on donnerait lieu à une absurdité:

$$(ab) + c = (a + c)\,(b + c).$$

En est-il toujours ainsi? N'existe-t-il pas d'algèbre telle que la loi distributive soit tout aussi symétrique par rapport à l'addition

et à la multiplication que les lois commutative et associative? Il se trouve que l'algèbre en question existe: c'est l'algèbre des ensembles, en laquelle toutes les trois lois sont symétriques par rapport aux opérations d'addition et de multiplication.

La ressemblance entre les opérations d'addition et de multiplication se manifeste également dans l'existence de deux nombres 0 et 1 remarquables dans ce sens que tout nombre additionné au premier et multiplié par le deuxième reste inchangé:

$$a + 0 = a, \quad a \cdot 1 = a.$$

Remarquons que la seconde relation se déduit de la première pour peu qu'on change (+) en (·) et 0 en 1.

Or, ici encore, l'analogie entre les opérations d'addition et de multiplication ne va pas trop loin. C'est ainsi que le nombre 0 joue un rôle assez particulier par rapport à tous les autres nombres, y compris l'unité. Ce rôle particulier du nombre 0 découle de la relation $a \cdot 0 = 0$. Remplaçant dans cette relation (·) par (+) et 0 par 1, on aboutit à la relation $a + 1 = 1$, qui ne se vérifie presque jamais.

On verra par la suite qu'en algèbre des ensembles la ressemblance entre le zéro et l'unité est sensiblement plus grande qu'en algèbre classique.

Ces remarques préliminaires faites, passons à l'étude des opérations effectuées sur les ensembles.

b) Réunion d'ensembles

On entend par réunion de deux ensembles X et Y l'ensemble constitué par les éléments et seulement par les éléments qui appartiennent au moins à l'un des ensembles X, Y, c.-à-d. soit à X, soit à Y. La réunion de X et de Y se désigne comme suit: $X \cup Y$. La définition formelle est

$$x \in X \cup Y \rightleftarrows x \in X \quad \text{ou} \quad x \in Y. \qquad (1\text{-}7)$$

On donne parfois à la réunion de deux ensembles le nom de *somme* de deux ensembles et on la désigne par $X + Y$. Or, les propriétés de la réunion sont quelque peu différentes de celles d'une somme arithmétique ordinaire. Aussi évitera-t-on l'usage de ce terme.

Exemple 1-1. Si $X = \{1, 2, 3, 4, 5\}$ et $Y = \{2, 4, 6, 7\}$, alors $X \cup Y = \{1, 2, 3, 4, 5, 6, 7\}$.

Exemple 1-2. Si X est l'ensemble des étudiants du groupe qui ont toutes les notes excellentes, et Y, l'ensemble des étudiants habitant le foyer, alors $X \cup Y$ est l'ensemble des étudiants qui ou bien ont toutes les notes excellentes, ou bien habitent le foyer.

Exemple 1-3. Considérons deux cercles (fig. 1-1). Si X est l'ensemble des points du cercle de gauche, et Y, celui du cercle de droite, alors $X \cup Y$ représente le domaine hachuré limité par les deux cercles.

La notion de réunion peut être étendue à un nombre plus grand d'ensembles. Désignons par $\mathfrak{M} = \{X_1, \ldots, X_n\}$ la collection de n ensembles $X_1, \ldots, X_n$, dite parfois *système* d'ensembles. La réunion de ces ensembles

$$\bigcup_{i=1}^{n} X_i = \bigcup_{X\in\mathfrak{M}}, \quad X = X_1 \cup \ldots \cup X_n \tag{1-8}$$

représente un ensemble, constitué par les éléments et seulement par les éléments qui appartiennent au moins à l'un des ensembles faisant partie du système $\mathfrak{M}$.

Les réunions d'ensembles vérifient les lois commutative et associative

$$X \cup Y = Y \cup X; \tag{1-9}$$

$$(X \cup Y) \cup Z = X \cup (Y \cup Z) = X \cup Y \cup Z, \tag{1-10}$$

ce qui découle du fait que les premiers et les deuxièmes membres des équations se composent des mêmes éléments. Ensuite,

$$X \cup \varnothing = X. \tag{1-11}$$

Fig. 1-1. Réunion de deux ensembles

Cette relation est aussi évidente, car un ensemble vide ne contient point d'éléments; donc, X et $X \cup \varnothing$ contiennent les mêmes éléments. On voit de (1-11) que l'ensemble vide $\varnothing$ joue le rôle de zéro en algèbre des ensembles. L'expression que l'on vient de considérer est analogue à l'expression $a + 0 = a$ propre à l'algèbre classique.

c) Intersection d'ensembles

On entend par *intersection* de deux ensembles X et Y l'ensemble constitué par les éléments et seulement par les éléments dont chacun appartient en même temps à X et à Y. L'intersection des ensembles X et Y se désigne par $X \cap Y$. La définition formelle est

$$x \in X \cap Y \rightleftarrows x \in X \text{ et } x \in Y. \tag{1-12}$$

Parfois, on désigne l'intersection de deux ensembles sous le terme de *produit* de deux ensembles et on la représente par le symbole XY. Or, les propriétés de l'intersection sont quelque peu différentes de celles d'un produit arithmétique ordinaire. Aussi évitera-t-on l'usage de ce terme.

Exemple 1-4. Pour les ensembles X et Y de l'exemple 1-1 on a $X \cap Y = \{2, 4\}$.

Exemple 1-5. Pour les ensembles X et Y de l'exemple 1-2 on a $X \cap Y$ pour désigner l'ensemble des étudiants du groupe n'ayant que des notes excellentes et habitant le foyer.

Exemple 1-6. Considérons deux cercles représentés sur la figure 1-2. Si X est l'ensemble des points du cercle de gauche et Y, celui du cercle de droite, l'expression $X \cap Y$ désigne le domaine hachuré qui représente la partie commune des deux cercles.

L'opération d'intersection permet d'établir plusieurs relations entre deux ensembles.

On dit que les ensembles X et Y sont *disjoints* s'ils n'ont aucun élément commun, en sorte que

$$X \cap Y = \varnothing. \tag{1-13}$$

Exemple 1-7. On considère comme ensembles disjoints:

1) les ensembles $\{1, 2, 3\}$ et $\{4, 5, 6\}$;

2) l'ensemble des étudiants du groupe qui ont toutes les notes excellentes et l'ensemble des étudiants qui ont les mauvaises notes;

3) les ensembles des points des cercles X et Y représentés sur la figure 1-3.

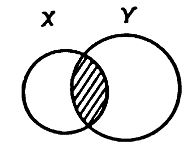


Fig. 1-2. Intersection de deux ensembles

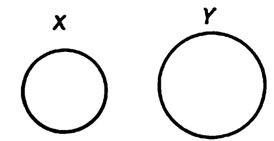


Fig. 1-3. Ensembles disjoints

On dit que les ensembles X et Y *se rencontrent* s'ils vérifient trois conditions:

il existe dans X un élément qui n'appartient pas à Y;

il existe dans Y un élément qui n'appartient pas à X;

il existe un élément appartenant aussi bien à X qu'à Y.

Soulignons une particularité de l'algèbre des ensembles relativement à celle des nombres. Deux nombres a et b peuvent être liés l'un à l'autre par trois relations, ou possibilités:

$$a < b, \quad a = b, \quad b < a. \tag{1-14}$$

Par contre, deux ensembles X et Y peuvent être tels que des trois relations

$$X \subset Y, \quad X = Y, \quad Y \subset X \tag{1-15}$$

aucune ne sera vérifiée. C'est ainsi que si X est l'ensemble des étudiants qui ont toutes les notes excellentes, et Y, l'ensemble des étudiants habitant le foyer, la signification des trois relations citées sera la suivante:

$X \subset Y$, chaque étudiant qui a toutes les notes excellentes habite obligatoirement le foyer;

$X = Y$, tous les étudiants qui ont toutes les notes excellentes, et seulement les étudiants de cette catégorie, habitent le foyer;

$Y \subset X$, tous les étudiants habitant le foyer ont toutes les notes excellentes.

De toute évidence, ces relations n'englobent pas toutes les possibilités. En réalité, comme il découle des définitions données plus haut, deux ensembles X et Y peuvent être liés l'un à l'autre par l'une des cinq relations suivantes:

$$X = Y; \quad X \subset Y; \quad Y \subset X; \quad X \cap Y = \varnothing;$$

X et Y se rencontrent.

La notion d'intersection peut être étendue à un nombre plus grand d'ensembles. Prenons un système d'ensembles $\mathfrak{M} = \{X_1, \ldots, X_n\}$. L'intersection de ces ensembles s'écrit sous la forme

$$\bigcap_{X \in \mathfrak{M}} X = \bigcap_{i=1}^{n} X_i = X_1 \cap \ldots \cap X_n \tag{1-16}$$

et représente un ensemble dont les éléments appartiennent à chacun des ensembles faisant partie de $\mathfrak{M}$.

Il est facile de voir que l'intersection d'ensembles est commutative

$$X \cap Y = Y \cap X \tag{1-17}$$

et associative

$$(X \cap Y) \cap Z = X \cap (Y \cap Z) = X \cap Y \cap Z. \tag{1-18}$$

Remarquons par ailleurs qu'il est légitime d'écrire

$$X \cap \varnothing = \varnothing, \tag{1-19}$$

relation analogue à la relation $a \cdot 0 = 0$ de l'algèbre classique. La relation (1-19), considérée conjointement avec la relation (1-11), montre que l'ensemble vide remplit la fonction de zéro en algèbre des ensembles.

Fig. 1-4. Différence de deux ensembles

d) Différence d'ensembles

Cette opération, à la différence de celles de réunion et d'intersection, ne vaut que pour deux ensembles. On entend par *différence* de deux ensembles X et Y l'ensemble constitué par les éléments et seulement par les éléments qui appartiennent à X et n'appartiennent pas à Y. Elle est désignée par $X \setminus Y$. Ainsi donc,

$$x \in X \setminus Y \rightleftarrows x \in X; \quad x \notin Y. \tag{1-20}$$

Exemple 1-8. Pour les ensembles X et Y de l'exemple 1-1 on a $X \setminus Y = \{1, 3, 5\}$, $Y \setminus X = \{6, 7\}$. Si X et Y sont les ensembles de l'exemple 1-2, alors $X \setminus Y$ est l'ensemble des étudiants qui ont toutes les notes excellentes et n'habitent pas le foyer. Pour X et Y de l'exemple 1-3, $X \setminus Y$ est le domaine hachuré de la figure 1-4.

e) Ensemble universel

Comme on vient de le voir, le rôle de zéro en algèbre des ensembles est joué par l'ensemble vide. Une question se pose de savoir s'il existe un ensemble I tel qu'il puisse remplir la fonction de l'unité, c.-à-d. vérifier la condition

$$X \cap I = X, \tag{1-21}$$

analogue à la condition $a \cdot 1 = a$ de l'algèbre classique.

La relation (1-21) veut dire que l'intersection, ou « partie commune », de l'ensemble I et de l'ensemble X pour tout ensemble X se confond avec cet ensemble lui-même. Or, cela n'est possible que si l'ensemble I contient tous les éléments constitutifs de l'ensemble X, de sorte que tout X est intégralement inclus dans I. L'ensemble I vérifiant cette condition s'appelle *achevé*, ou *universel*, ou *unitaire*.

Compte tenu de ce qui précède, on définit l'ensemble universel comme suit. Si une considération donnée ne porte que sur des sous-ensembles d'un certain ensemble fixé I, ce dernier ensemble I, le plus grand de tous, porte le nom d'ensemble universel.

Il est à noter que le rôle de l'ensemble universel peut être joué par des ensembles différents, selon le cas. Ainsi, quand on considère divers ensembles des étudiants du groupe (étudiants qui ont toutes les notes excellentes; étudiants qui touchent la bourse; étudiants qui habitent le foyer, etc.), c'est l'ensemble des étudiants du groupe qui se présente comme ensemble universel.

Il est commode de représenter graphiquement l'ensemble universel sous la forme de l'ensemble des points d'un rectangle. Certains domaines à l'intérieur de ce rectangle représenteront alors divers sous-ensembles de l'ensemble universel. Le dessin représentant les ensembles sous la forme de domaines à l'intérieur d'un rectangle assimilé à l'ensemble universel est connu sous le nom de *diagramme d'Euler-Venn*.

L'ensemble universel possède une propriété remarquable, qui n'a pas son analogue en algèbre classique, à savoir, pour tout ensemble X on a la relation

$$X \cup I = I. \tag{1-22}$$

En effet, la réunion $X \cup I$ est un ensemble contenant aussi bien tous les éléments de X que tous les éléments de I. Or, l'ensemble I contient a priori tous les éléments de X, de sorte que $X \cup I$ contient les mêmes éléments que I et représente donc l'ensemble universel I lui-même.

f) Complémentaire d'un ensemble

Un ensemble $\overline{X}$ défini par la relation

$$\overline{X} = I \setminus X \qquad (1\text{-}23)$$

est dit *complémentaire* de l'ensemble X par rapport à (ou dans) l'ensemble universel I. Sur le diagramme de la figure 1-5 l'ensemble $\overline{X}$ représente la partie non hachurée. La définition formelle est:

$$\overline{X} = \{x : x \in I \text{ et } x \notin X\}.$$

Exemple 1-9. Si $I = \{1, 2, 3, 4, 5, 6, 7\}$ et $X = \{3, 5, 7\}$, alors $\overline{X} = \{1, 2, 4, 6\}$.

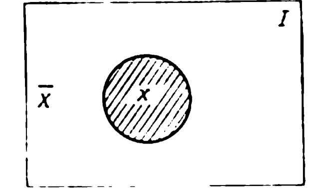


Fig. 1-5. Complémentaire d'un ensemble

Il vient de (1-23) que X et $\overline{X}$ n'ont pas d'élément commun, de sorte que

$$X \cap \overline{X} = \varnothing. \qquad (1\text{-}24)$$

En outre, il n'existe pas d'élément de I tel qu'il n'appartienne soit à X, soit à $\overline{X}$, car les éléments n'appartenant pas à X appartiennent forcément à $\overline{X}$. Donc,

$$X \cup \overline{X} = I. \qquad (1\text{-}25)$$

La symétrie de la formule (1-25) par rapport à X et $\overline{X}$ signifie non seulement que $\overline{X}$ est complémentaire de X mais aussi que X l'est de $\overline{X}$. Or, le complémentaire de $\overline{X}$ est $\overline{\overline{X}}$. On a donc

$$\overline{\overline{X}} = X. \qquad (1\text{-}26)$$

A l'aide du complémentaire, on réussit à représenter sous une forme commode la différence de deux ensembles

$$X \setminus Y = \{x : x \in X \text{ et } x \notin Y\} = \{x : x \in X \text{ et } x \in \overline{Y}\},$$

c.-à-d.

$$X \setminus Y = X \cap \overline{Y}. \qquad (1\text{-}27)$$

g) Partition d'un ensemble

Une des opérations effectuées très souvent sur les ensembles est la *partition* d'un ensemble en un système de sous-ensembles. Par exemple, le système de promotions dans le cadre de la faculté donnée est une partition de l'ensemble des étudiants de la faculté ; le système de groupes de la promotion est une partition de l'ensemble des étudiants de la promotion. Si N est l'ensemble des nombres naturels

et A_0, A_1, les ensembles des nombres pairs et impairs, le système $\{A_0, A_1\}$ constitue une partition de N. On peut partager l'ensemble des nombres naturels d'une autre façon, par exemple en ensembles des nombres divisibles par 3 sans reste; avec reste 1; avec reste 2. La production d'une entreprise est partagée en un système d'ensembles constitués par les produits de première qualité, ceux de seconde qualité et le rebut. Le nombre d'exemples analogues pourrait être multiplié à l'infini.

Pour donner à la notion de partition une définition rigoureuse, considérons un ensemble M et un système d'ensembles $\mathfrak{M} = \{X_1, \ldots, X_n\}$. On dit que le système d'ensembles $\mathfrak{M}$ est la *partition* de l'ensemble M s'il satisfait aux conditions suivantes:

1) tout ensemble X de $\mathfrak{M}$ est sous-ensemble de M:

$$\forall X \in \mathfrak{M}: X \subseteq M; \tag{1-28}$$

2) deux ensembles quelconques X et Y de $\mathfrak{M}$ sont disjoints:

$$\forall X \in \mathfrak{M}, \quad \forall Y \in \mathfrak{M}: X \neq Y \rightarrow X \cap Y = \varnothing; \tag{1-29}$$

3) la réunion de tous les ensembles faisant partie de la partition forme l'ensemble M:

$$\bigcup_{X \in \mathfrak{M}} X = M. \tag{1-30}$$

Nous reparlerons de la partition à propos de la relation d'équivalence, qui est étroitement liée à la première.

h) Identités en algèbre des ensembles

Les opérations de réunion, d'intersection et la notion de complémentaire permettent d'écrire différentes expressions algébriques à partir d'ensembles. Désignons par $\mathfrak{A}(X, Y, Z)$ une expression algébrique composée par les ensembles X, Y et Z. Cette expression est elle-même un ensemble. Soit $\mathfrak{B}(X, Y, Z)$ une autre expression algébrique composée par ces mêmes ensembles. Si les deux expressions algébriques représentent un seul et même ensemble, on peut les identifier: nous obtenons alors une identité algébrique de la forme

$$\mathfrak{A}(X, Y, Z) = \mathfrak{B}(X, Y, Z). \tag{1-31}$$

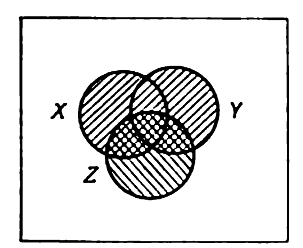
De tels ensembles sont d'une grande utilité lorsqu'il s'agit de transformer des expressions algébriques écrites au moyen d'ensembles; nous allons considérer quelques ensembles de ce type.

1. Sur la figure 1-6 sont représentés les diagrammes d'Euler et Venn pour les expressions $(X \cup Y) \cap Z$ et $(X \cap Z) \cup (Y \cap Z)$. On voit sur ces diagrammes que les deux expressions définissent un seul et

même ensemble, de sorte qu'en algèbre des ensembles a lieu l'identité

$$(X \cup Y) \cap Z = (X \cap Z) \cup (Y \cap Z), \qquad (1\text{-}32)$$

analogue à la loi distributive $(a + b)\,c = ac + bc$ propre à l'algèbre classique.



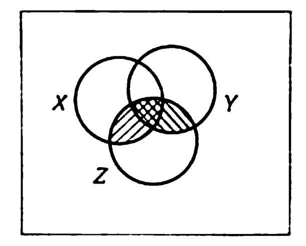
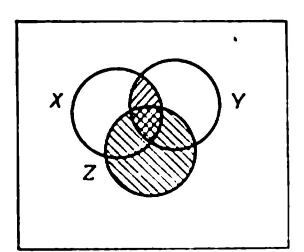


Fig. 1-6. Illustration géométrique de l'identité $(X \cup Y) \cap Z = (X \cap Z) \cup \cup (Y \cap Z)$

2. Nous ne pouvons, en algèbre classique, remplacer dans la loi distributive l'addition par la multiplication et vice versa, au risque d'obtenir une expression absurde $(ab) + c = (a + c)\,(b + c)$. Il en est tout autrement en algèbre des ensembles.



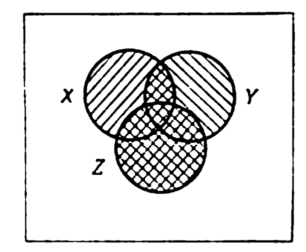


Fig. 1-7. Illustration géométrique de l'identité $(X \cap Y) \cup Z = (X \cup Z) \cap (Y \cup Z)$

On voit sur la figure 1-7 les diagrammes d'Euler et Venn pour les expressions algébriques $(X \cap Y) \cup Z$ et $(X \cup Z) \cap (Y \cup Z)$. Les deux expressions conduisent à un seul et même ensemble. On a donc l'identité

$$(X \cap Y) \cup Z = (X \cup Z) \cap (Y \cup Z). \qquad (1\text{-}33)$$

3. On vérifie facilement que si $Y \subseteq X$, alors

$$X \cap Y = Y, \quad X \cup Y = X. \qquad (1\text{-}34)$$

En effet, tous les éléments de l'ensemble Y sont en même temps éléments de l'ensemble X. Donc, l'intersection de ces ensembles, c.-à-d. la partie commune de X et Y, se confond avec Y. La contri-

bution apportée par l'ensemble Y à la réunion des ensembles X et Y ne contient pas un seul élément nouveau, celui-ci étant déjà, forcément, élément de X. Par conséquent, $X \cup Y$ se confond avec X.

4. Posant dans (1-34) $Y = X$ et se rappelant que $X \subseteq X$, on trouve:

$$X \cap X = X, \quad X \cup X = X. \tag{1-35}$$

La démonstration des identités en algèbre des ensembles au moyen des diagrammes d'Euler et Venn s'avère quelquefois incommode. Il y a un moyen plus général de démonstration de l'identité de deux expressions algébriques.

Désignons, comme auparavant, par $\mathfrak{A}(X, Y, Z)$ et $\mathfrak{B}(X, Y, Z)$ deux expressions algébriques formées à partir des ensembles X, Y et Z par application de la réunion, de l'intersection et du complémentaire. Pour démontrer que $\mathfrak{A} = \mathfrak{B}$, il suffit de montrer que $\mathfrak{A} \subseteq \mathfrak{B}$ et que $\mathfrak{B} \subseteq \mathfrak{A}$. D'autre part, pour montrer que $\mathfrak{A} \subseteq \mathfrak{B}$, il faut que $x \in \mathfrak{A}$ implique $x \in \mathfrak{B}$. De façon analogue, pour démontrer que $\mathfrak{B} \subseteq \mathfrak{A}$, il faut s'assurer que $x \in \mathfrak{B}$ implique $x \in \mathfrak{A}$. Procédons par cette méthode et démontrons encore quelques identités.

5. Démontrons l'identité

$$\overline{X \cup Y} = \overline{X} \cap \overline{Y}. \tag{1-36}$$

Supposons que $x \in \overline{X \cup Y}$, c.-à-d. que $x \notin X \cup Y$. Cela revient à dire que $x \notin X$ et $x \notin Y$, c.-à-d. que $x \in \overline{X}$ et $x \in \overline{Y}$. Par conséquent, $x \in \overline{X} \cap \overline{Y}$. Supposons maintenant que $y \in \overline{X} \cap \overline{Y}$, c.-à-d. que $y \in \overline{X}$ et $y \in \overline{Y}$. Cela signifie que $y \notin X$ et $y \notin Y$, c.-à-d. que $y \notin X \cup Y$. Donc, $y \in \overline{X \cup Y}$.

6. Démontrons l'identité

$$\overline{X \cap Y} = \overline{X} \cup \overline{Y} \tag{1-37}$$

en mettant ses deux membres sous une forme identique. Cherchant les complémentaires des deux membres de (1-37), on obtient $\overline{\overline{X \cap Y}} = \overline{\overline{X} \cup \overline{Y}}$. Le premier membre devient $X \cap Y$. On obtient la même expression du second membre en le transformant d'après (1-36).

Dans les ouvrages mathématiques, les identités (1-36) et (1-37) sont appelées d'habitude *identités de Morgan*.

1-3. MISE EN ORDRE DES ÉLÉMENTS. PRODUIT DIRECT D'ENSEMBLES

a) Ensemble ordonné

A côté de la notion d'ensemble comme collection d'éléments, il existe une notion fort importante d'ensemble ordonné, ou cortège. On entend par *cortège* une suite d'éléments, c.-à-d. une collec-

tion d'éléments organisée de telle façon que chaque élément y occupe une place déterminée. Les éléments eux-mêmes sont désignés alors sous le terme de *composantes* du cortège: première composante, deuxième composante, et ainsi de suite. Voici quelques exemples de cortèges: l'ensemble des personnes formant une file d'attente, ou queue; l'ensemble des mots d'une phrase; les nombres définissant la longitude et la latitude d'un point sur le terrain, etc. Dans tous ces ensembles la place de chaque élément est parfaitement déterminée et ne peut être changée arbitrairement.

Le nombre des éléments formant un cortège est la *longueur* de celui-ci. Pour désigner un cortège, nous utiliserons les parenthèses. Ainsi, l'ensemble

$$a = (a_1, a_2, \ldots, a_n) \tag{1-38}$$

est un cortège de longueur n et d'éléments $a_1, \ldots, a_n$. Les cortèges de longueur 2 sont appelés *couples* ou *couples ordonnés*; ceux de longueur 3, *triplets*; ceux de longueur 4, *quadruplets*, etc. Dans le cas général, un cortège de longueur n est un *n-uplet*. Des cas particuliers d'un cortège sont le cortège (a) de longueur 1 et le cortège vide de longueur 0, désigné par ( ) ou par Λ. A l'opposé d'un ensemble ordinaire, le cortège peut contenir des éléments identiques: deux mots identiques dans la phrase, mêmes valeurs numériques de longitude et de latitude d'un point sur le terrain, etc.

Dans le texte qui suit, nous allons considérer des ensembles ordonnés qui ont pour éléments des nombres réels. Les ensembles ordonnés de ce type sont appelés *points de l'espace* ou *vecteurs*. C'est ainsi que le cortège (a_1, a_2) peut être considéré comme un point sur le plan, ou bien comme un vecteur allant de l'origine des coordonnées au point donné (fig. 1-8,*a*). Les composantes a_1 et a_2 deviendront les projections du vecteur sur les axes *1* et *2*

$$\mathrm{pr}_1 (a_1, a_2) = a_1; \quad \mathrm{pr}_2 (a_1, a_2) = a_2.$$

Le cortège (a_1, a_2, a_3) peut être considéré comme un point dans l'espace à trois dimensions, ou bien comme un vecteur tridimensionnel reliant l'origine des coordonnées au point en question (fig. 1-8,*b*). Les projections du vecteur sur les axes de coordonnées sont

$$\mathrm{pr}_i (a_1, a_2, a_3) = a_i, \quad i = 1, 2, 3.$$

Cependant, dans le cas donné, on peut parler de la projection du cortège simultanément sur deux axes, disons *1* et *2*, c.-à-d. sur le plan de coordonnées. Il est facile de voir que cette projection représente un cortège à deux éléments

$$\mathrm{pr}_{12} (a_1, a_2, a_3) = (a_1, a_2).$$

Généralisant ces notions, assimilons un ensemble ordonné de nombres réels à n éléments $(a_1, \ldots, a_n)$ à un point dans un espace imaginaire à n dimensions (dit parfois *hyperespace*), ou bien à un

vecteur à n dimensions. Les composantes du cortège a à n éléments seront considérées alors comme les projections de ce cortège sur les axes correspondants :

$$\mathrm{pr}_i a = a_i, \quad i = 1, \ldots, n. \tag{1-39}$$

Numérotant les axes par $i, j, \ldots, l$ de façon que $1 \leqslant i < j < \ldots$

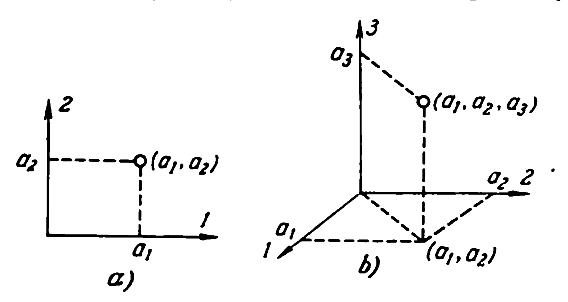


Fig. 1-8. Projections de cortèges à deux et à trois éléments

$\ldots < l \leqslant n$, on obtient la projection du cortège a sur les axes ., $j, \ldots, l$ sous la forme

$$\mathrm{pr}_{i,\, j,\, \ldots,\, l}\, a = (a_i,\ a_j,\ \ldots,\ a_l). \tag{1-40}$$

La projection du cortège sur un ensemble vide des axes est le cortège vide

$$\mathrm{pr}_{\varnothing} a = \Lambda. \tag{1-41}$$

C'est au chapitre 3 que l'on trouve une définition plus complète et plus rigoureuse de l'espace multidimensionnel.

b) Produit direct d'ensembles

On entend par *produit direct* de deux ensembles X et Y l'ensemble (noté $X \times Y$) constitué entièrement et seulement par les couples ordonnés dont la première composante appartient à X et la deuxième à Y. Ainsi donc, les éléments de l'ensemble ordonné sont des cortèges à deux éléments de la forme (x, y). La définition formelle est

$$X \times Y = \{(x, y) : x \in X, \quad y \in Y\}. \tag{1-42}$$

Exemple 1-10. Soient $X = \{1, 2\}$, $Y = \{1, 3, 4\}$. Alors $X \times Y = \{(1,1), (1,3), (1,4), (2,1), (2,3), (2,4)\}$. La représentation géométrique de cet ensemble est donnée sur la figure 1-9, *a*.

Exemple 1-11. Soient X et Y des segments de l'axe réel. On représente le produit direct $X \times Y$ par un rectangle (hachuré sur la figure 1-9, *b*). On peut voir du dessin que les propriétés du produit direct diffèrent de celles d'un produit arithmétique ordinaire. En particulier, le produit direct est sensible à la permutation de ses facteurs :

$$X \times Y \neq Y \times X. \tag{1-43}$$

La notion de produit direct s'étend sans difficulté à un plus grand nombre d'ensembles. On entend par produit direct des ensembles $X_1, X_2, \ldots, X_r$ l'ensemble noté $X_1 \times X_2 \times \ldots \times X_r$ et cons-

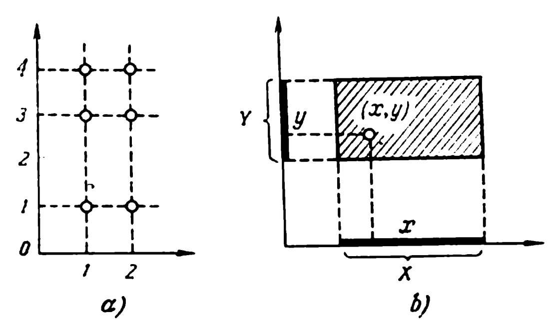


Fig. 1-9. Illustration géométrique du produit direct de deux ensembles

titué entièrement et seulement par les cortèges de longueur r, dont la première composante appartient à X_1, la deuxième à X_2, et ainsi de suite.

On s'assure facilement que

$$X \times Y = \varnothing \rightleftarrows X = \varnothing \quad \text{ou} \quad Y = \varnothing, \qquad (1\text{-}44)$$

puisqu'il n'y a pas de couple ordonné amputé de sa première ou de sa deuxième composante. Par analogie, $X_1 \times X_2 \times \ldots \times X_r = \varnothing$ si et seulement si au moins un des ensembles $X_1, X_2, \ldots, X_r$ est vide.

Un cas particulier du produit direct est la *puissance* d'un ensemble. Soit donné un ensemble quelconque M. On désigne par s-ième puissance de M (notation M^s) le produit direct de s ensembles identiques égaux à M:

$$M^s = \underbrace{M \times M \times \ldots \times M}_{s \text{ fois}}. \qquad (1\text{-}45)$$

Cette définition vaut pour $s = 2, 3, \ldots$ On l'étend à un s entier non négatif quelconque en spécifiant que

$$M^1 = M, \; M^0 = \{\Lambda\}. \qquad (1\text{-}46)$$

Si R est l'ensemble des nombres réels, alors $R^2 = R \times R$ est le plan réel, et $R^3 = R \times R \times R$ l'espace réel tridimensionnel.

c) Projection d'un ensemble

La projection d'un ensemble, étroitement liée à celle d'un cortège, n'existe que pour les ensembles dont les éléments sont constitués par des cortèges de même longueur.

Soit M un ensemble constitué par des cortèges de longueur s. Sa projection sera représentée alors par l'ensemble des projections des cortèges de M.

Exemple 1-12. Soit $M = \{(1, 2, 3, 4, 5), (2, 1, 3, 5, 5), (3, 3, 3, 3, 3), (3, 2, 3, 4, 3)\}$. Alors

$$\mathrm{pr}_2 M = \{2, 1, 3\}; \quad \mathrm{pr}_{2,4} M = \{(2, 4), (1, 5), (3, 3)\}.$$

On vérifie sans peine que si $M = X \times Y$, alors

$$\mathrm{pr}_1 M = X; \quad \mathrm{pr}_2 M = Y, \tag{1-47}$$

et si $Q \subseteq X \times Y$, alors

$$\mathrm{pr}_1 Q \subseteq X; \quad \mathrm{pr}_2 Q \subseteq Y. \tag{1-48}$$

1-4. CORRESPONDANCES

a) Définition d'une correspondance

Considérons deux ensembles X et Y. On peut faire correspondre les éléments de ces ensembles, d'une façon ou d'une autre, les uns aux autres, en sorte qu'ils forment des couples (x, y). Si le mode de la correspondance est défini, c.-à-d. si l'on connaît pour tout $x \in X$ l'élément $y \in Y$ auquel on fait correspondre x, on dit qu'il y a *correspondance* entre X et Y [8, 11]. Remarquons qu'il n'est absolument pas indispensable que la correspondance porte sur la totalité des éléments des ensembles X et Y.

Pour définir une correspondance, il faut que l'on connaisse:

1) l'ensemble X dont les éléments sont censés correspondre à ceux d'un autre ensemble;

2) l'ensemble Y aux éléments duquel on fait correspondre les éléments d'un premier ensemble;

3) l'ensemble $Q \subseteq X \times Y$ définissant la loi conformément à laquelle se réalise la correspondance, c.-à-d. l'ensemble énumérant tous les couples (x, y) faisant partie de la correspondance. De cette façon, la correspondance, désignée par q, représente un triplet d'ensembles

$$q = (X, Y, Q) \tag{1-49}$$

dans lequel $Q \subseteq X \times Y$. Dans cette expression la première composante X est dite *ensemble* (ou *domaine*) *de départ*, la deuxième composante Y *ensemble d'arrivée*, la troisième composante Q *graphique* de la correspondance. Le terme « graphique » sera expliqué plus en détail au cours de l'étude d'une forme particulière de la correspondance qui porte le nom de fonction.

En plus des trois ensembles considérés X, Y et Q, à chaque correspondance sont étroitement liés encore deux ensembles, à savoir: l'ensemble $\mathrm{pr}_1 Q$, dit *ensemble de définition*, qui comprend les éléments de X faisant partie de la correspondance, et l'ensemble $\mathrm{pr}_2 Q$, dit

ensemble de valeurs, qui réunit les éléments de Y faisant partie de la correspondance.

Si $(x, y) \in Q$, on dit que y correspond à x. La correspondance se représente géométriquement, d'une façon très commode, par une flèche allant de x vers y.

Exemple 1-13. Soient $X = \{1,2\}$, $Y = \{3,5\}$, de sorte que $X \times Y = \{(1,3), (1,5), (2,3), (2,5)\}$. Cet ensemble donne lieu à 16 correspondances différentes. Citons-en quelques-unes:

$$Q_1 = \{(1,3)\};\ \mathrm{pr}_1 Q_1 = \{1\};\ \mathrm{pr}_2 Q_1 = \{3\};$$
$$Q_2 \{(1,3), (1,5)\};\ \mathrm{pr}_1 Q_2 = \{1\};\ \mathrm{pr}_2 Q_2 = \{3,5\} = Y.$$

Exemple 1-14. L'entreprise dispose de trois véhicules: deux camions α et β utilisés en deux relèves et un autocar γ qui est rarement utilisé. Le camion β est en réparation. L'entreprise emploie trois chauffeurs: a, b et c, dont c est en vacances. L'affectation des chauffeurs aux véhicules est une correspondance. Une des correspondances possibles sera:

$$q = (\{a, b, c\}, \{\alpha, \beta, \gamma\}, \{(a, \alpha), (a, \gamma), (b, \alpha)\}).$$

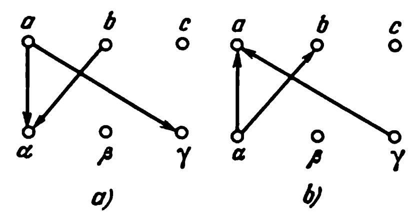


Fig. 1-10. Représentation géométrique des correspondances directe et inverse

La représentation géométrique de cette correspondance est donnée sur la figure 1-10, *a*. L'élément α y correspond aux éléments a et b, tandis que l'élément γ correspond à a. La correspondance q est définie sur a et b mais non sur c; donc, l'ensemble de définition est $\{a, b\}$. L'ensemble de valeurs est $\{\alpha, \gamma\}$.

b) Correspondance inverse

Pour toute correspondance $q = (X, Y, Q)$, $Q \subseteq X \times Y$ il existe la *correspondance inverse* que l'on obtient en considérant la correspondance donnée dans le sens inverse, c.-à-d. en déterminant les éléments $x \in X$ auxquels on fait correspondre les éléments $y \in Y$. La correspondance inverse de q sera désignée par

$$q^{-1} = (Y, X, Q^{-1}), \tag{1-50}$$

où $Q^{-1} \subseteq Y \times X$.

Exemple 1-15. Dans les conditions de l'exemple 1-14 la correspondance inverse est l'affectation des véhicules aux chauffeurs

$$(\{\alpha, \beta, \gamma\} \{a, b, c\}, \{(\alpha, a), (\alpha, b), (\gamma, a)\})$$

représentée géométriquement sur la figure 1-10, *b*.

L'exemple considéré laisse voir que l'on obtient la représentation géométrique de la correspondance inverse en changeant la direction des flèches dans la représentation géométrique de la correspondance

directe. Il s'ensuit qu'une correspondance inverse a pour son inverse la correspondance directe:

$$(q^{-1})^{-1} = q. \qquad (1\text{-}51)$$

c) Composée des correspondances

On entend par *composée des correspondances*, deux correspondances écrites l'une après l'autre.

La composée des correspondances est le résultat d'une opération effectuée sur trois ensembles X, Y et Z sur lesquels sont définies deux correspondances

$$\left.\begin{array}{ll} q = (X, Y, Q), & Q \subseteq X \times Y; \\ p = (Y, Z, P), & P \subseteq Y \times Z, \end{array}\right\} \qquad (1\text{-}52)$$

l'ensemble de valeurs de la première correspondance coïncidant avec l'ensemble de définition de la seconde

$$\mathrm{pr}_2 Q = \mathrm{pr}_1 P. \qquad (1\text{-}53)$$

La première correspondance définit pour tout $x \in \mathrm{pr}_1 Q$ un élément, peut-être non unique, $y \in Y$. Conformément à la définition de la composée des correspondances, on doit maintenant définir pour $y \in Y$ trouvé, en employant la seconde correspondance, l'élément $z \in Z$. Ainsi donc, la composée des correspondances fait correspondre à chaque élément x de l'ensemble de définition de la première correspondance $\mathrm{pr}_1 Q$ un ou plusieurs éléments z de l'ensemble de valeurs de la seconde correspondance $\mathrm{pr}_2 P$.

Désignons la composée des correspondances q et p par $q(p)$, et le graphique de la composée des correspondances, par $Q \circ P$. La composée des correspondances (1-52) s'écrit alors sous la forme

$$q(p) = (X, Z, Q \circ P), \quad Q \circ P \subseteq X \times Z. \qquad (1\text{-}54)$$

Exemple 1-16. Si q est la correspondance définissant l'affectation des chauffeurs aux véhicules et p celle définissant l'affectation des véhicules aux itinéraires, la correspondance $q(p)$ définira l'affectation des chauffeurs aux itinéraires.

Il est naturel que la composée s'étende sans difficulté à plus de deux correspondances.

1-5. APPLICATIONS ET FONCTIONS

a) Applications et leurs propriétés

Soient donnés deux ensembles X et Y et, d'autre part, $\Gamma \subseteq X \times Y$ de sorte que $\mathrm{pr}_1 \Gamma = X$. Le triplet d'ensembles (X, Y, Γ) définit une certaine correspondance, remarquable par ailleurs par le fait que

son ensemble de définition $\mathrm{pr}_1\Gamma$ coïncide avec l'ensemble de départ, c.-à-d. avec X et que, par conséquent, la correspondance en question est partout définie sur X. En d'autres termes, pour tout $x \in X$ il existe un $y \in Y$ tel que $(x, y) \in \Gamma$. Une telle correspondance partout définie porte le nom d'*application* de X dans Y et s'écrit

$$\Gamma : X \to Y. \tag{1-55}$$

Très souvent, en employant le mot « application », on sous-entend une application univoque. Ce ne sera pas pourtant notre principe: nous admettrons qu'une application Γ fait correspondre à tout élément $x \in X$ un certain sous-ensemble

$$\Gamma x \subseteq Y, \tag{1-56}$$

appelé image de x. La loi régissant la correspondance est définie par l'ensemble Γ.

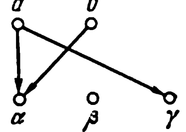


Fig. 1-11. Représentation géométrique de l'application

Exemple 1-17. Si dans l'exemple 1-14 on ne prend pas en considération le chauffeur c, on obtient l'application $\Gamma: X \to Y$ dans laquelle $X = \{a, b\}$ est l'ensemble des chauffeurs; $Y = \{\alpha, \beta, \gamma\}$ l'ensemble des véhicules; $\Gamma = \{(a, \alpha), (a, \gamma), (b, \alpha)\}$ l'affectation des chauffeurs aux véhicules. La représentation géométrique de cette application est donnée sur la figure 1-11.

Considérons quelques propriétés de l'application. Soit $A \subseteq X$. Pour tout $x \in A$ l'image de x sera l'ensemble $\Gamma x \subseteq Y$. La totalité des éléments de Y, images Γx pour tous les $x \in A$, sera appelée *image de l'ensemble* A et notée ΓA. En vertu de cette définition

$$\Gamma A = \bigcup_{x \in A} \Gamma x. \tag{1-57}$$

Si A_1 et A_2 sont sous-ensembles de X, on a

$$\Gamma (A_1 \cup A_2) = \Gamma A_1 \cup \Gamma A_2. \tag{1-58}$$

En effet,

$$\Gamma (A_1 \cup A_2) = \bigcup_{x \in A_1 \cup A_2} \Gamma x = (\bigcup_{x \in A_1} \Gamma x) \cup (\bigcup_{x \in A_2} \Gamma x) = \Gamma A_1 \cup \Gamma A_2.$$

Cependant, la relation

$$\Gamma (A_1 \cap A_2) = \Gamma A_1 \cap \Gamma A_2 \tag{1-59}$$

ne se vérifie que si l'application est univoque. Pour le démontrer, procédons à la partition des ensembles A_1 et A_2 (fig. 1-12) de la forme

$$A_1 = X_1 \cup X_0, \quad A_2 = X_2 \cup X_0,$$

où $X_0 = A_1 \cap A_2$.

Les ensembles X_1, X_2, X_0 representent alors une partition de l'ensemble $A_1 \cup A_2$. Bien que ces ensembles n'aient pas d'éléments communs, leurs images peuvent en comporter quand l'application

est non univoque. Donc,

$$\Gamma A_1 \cap \Gamma A_2 = \Gamma (X_1 \cup X_0) \cap \Gamma (X_2 \cup X_0) =$$
$$= (\Gamma X_1 \cap \Gamma X_2) \cup (\Gamma X_1 \cap \Gamma X_0) \cup (\Gamma X_2 \cap \Gamma X_0) \cup \Gamma X_0.$$

De cette relation on voit que (1-59) ne se réalise que lorsque

$$\Gamma X_i \cap \Gamma X_k = \varnothing; \quad i, k \in \{0, 1, 2\}; \quad i \neq k, \tag{1-60}$$

c.-à-d. quand l'application est univoque. Par contre, dans le cas général on a

$$\Gamma X_0 = \Gamma (A_1 \cap A_2) \subseteq \Gamma A_1 \cap \Gamma A_2. \tag{1-61}$$

Ces relations s'étendent facilement à un plus grand nombre de sous-ensembles A_i. Par exemple, si $A_1, \ldots, A_n$ sont sous-ensembles de X, alors

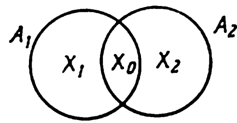


Fig. 1-12. Partition de deux ensembles non disjoints

$$\Gamma \left(\bigcup_{i=1}^{n} A_i \right) = \bigcup_{i=1}^{n} \Gamma A_i; \tag{1-62}$$

$$\Gamma \left(\bigcap_{i=1}^{n} A_i \right) \subseteq \bigcap_{i=1}^{n} \Gamma A_i. \tag{1-63}$$

N'étant qu'un cas particulier de la correspondance, l'application admet aussi bien les notions d'application inverse et de composée, notions qui ont été introduites à l'étude des correspondances.

b) Applications définies sur un seul ensemble

Un cas particulier très important de l'application a lieu quand les ensembles X et Y se confondent. L'application $\Gamma: X \to X$ représente alors l'application de X dans lui-même et se définit par le couple

$$(X, \Gamma), \tag{1-64}$$

où $\Gamma \subseteq X^2$. De pareilles applications sont étudiées en théorie des graphes; les éléments de cette théorie sont exposés au chapitre 2. Dans ce paragraphe, nous parlerons seulement de quelques opérations effectuées sur ces applications.

Soient Γ et Δ applications de l'ensemble X dans X. La composée de ces applications sera l'application $\Gamma\Delta$ qui, en vertu de la règle énoncée au § 1-4, se définit de la manière suivante:

$$(\Gamma\Delta)\, x = \Gamma\, (\Delta x). \tag{1-65}$$

Dans le cas particulier où $\Delta = \Gamma$, on obtient les applications

$$\Gamma^2 x = \Gamma\, (\Gamma x); \tag{1-66}$$
$$\Gamma^3 x = \Gamma\, (\Gamma^2 x), \text{ etc.} \tag{1-67}$$

Ainsi donc, dans le cas général pour tout $s \geqslant 2$ on a

$$\Gamma^s x = \Gamma(\Gamma^{s-1} x). \tag{1-68}$$

Par définition spéciale, introduisons la relation

$$\Gamma^0 x = x. \tag{1-69}$$

Il devient possible d'étendre la relation (1-68) à des s négatifs. En effet, on a en vertu de (1-68)

$$\Gamma^0 x = \Gamma(\Gamma^{-1} x) = \Gamma\Gamma^{-1} x = x. \tag{1-70}$$

Cela signifie que $\Gamma^{-1}x$ est une application inverse. Alors

$$\Gamma^{-2} x = \Gamma^{-1}(\Gamma^{-1} x), \tag{1-71}$$

et ainsi de suite.

Exemple 1-18. Soit X l'ensemble des hommes. Pour tout homme $x \in X$, désignons par Γx l'ensemble de ses enfants. Alors $\Gamma^2 x$ sera l'ensemble des petits-enfants de x, $\Gamma^3 x$ l'ensemble de ses arrière-petits-enfants, $\Gamma^{-1} x$ l'ensemble de ses parents, etc.

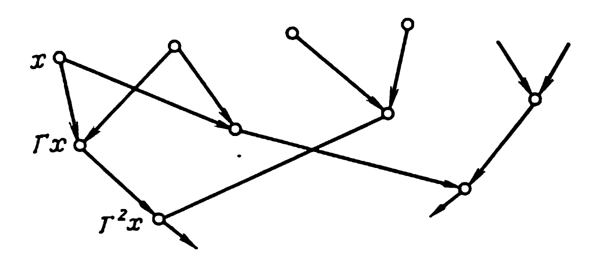


Fig. 1-13. Arbre généalogique

Représentant les hommes par des points et tirant des flèches allant de x vers Γx, on obtient l'arbre généalogique, tel qu'on le voit, par exemple, à la figure 1-13.

Exemple 1-19. Dans un jeu d'échecs, désignons par x une certaine position (c.-à-d. la disposition des pièces sur l'échiquier) qui peut avoir lieu au cours du jeu, et par X l'ensemble des positions possibles. Alors, pour tout $x \in X$, on représentera par Γx l'ensemble des positions qui découlent de x après un coup fait en conformité avec les règles du jeu. Dans ce cas

$\Gamma x = \varnothing$ si x est un mat ou un pat;

$\Gamma^3 x$ est l'ensemble des positions que l'on obtient de x en trois coups;

$\Gamma^{-1} x$ est l'ensemble des positions qui se ramènent à la position donnée en un seul coup.

Pour les applications définies sur un seul ensemble, on emploie aussi d'autres appellations que l'on rencontrera par la suite.

Par exemple, si les éléments $x \in X$ représentent les états d'un système dynamique, on peut considérer l'application Γx comme l'ensemble des états du système postérieurs à l'état donné. Il est

commode d'employer en cette occurrence le terme de *transformation* de l'état du système dynamique. Pour désigner certaines formes particulières d'applications définies sur un seul et même ensemble, on emploie aussi le terme de *relation*.

c) Fonction, fonctionnelle, opérateur

Considérons une application

$$f : X \rightarrow Y. \tag{1-72}$$

Cette application reçoit le nom de *fonction* si elle est univoque, c.-à-d. si pour tout couple $(x_1, y_1) \in f$ et $(x_2, y_2) \in f$ l'égalité $x_2 = x_1$ implique $y_2 = y_1$.

De la définition de l'application et des exemples qui viennent d'être étudiés, il est clair que les éléments des ensembles X et Y peuvent être les objets d'un caractère tout à fait arbitraire. Or, dans les problèmes de cybernétique, on attache un intérêt particulier aux applications qui sont univoques et dont l'ensemble de valeurs représente l'ensemble des nombres réels R. L'application univoque f définie par (1-72) est appelée *fonction à valeurs réelles* si $Y \subseteq R$.

Exemple 1-20. Supposons qu'on puisse aller de la ville donnée à une autre ville par train, par autocar ou par avion. Le prix du billet est respectivement de 7, 9 et 12 roubles. Dans cet exemple, le prix du billet peut être considéré comme une fonction du moyen de transport. Pour ce faire, envisageons les ensembles

$$X = \{\text{tr., car, av.}\},\ Y = \{7, 9, 12\}.$$

La fonction $f : X \rightarrow Y$ obtenue dans les conditions de l'exemple donné peut être mise sous la forme de l'ensemble $f = \{(\text{tr.}, 7), (\text{car}, 9), (\text{av.}, 12)\}$.

Dans chacun des couples $(x, y) \in f$ la valeur de y est dite fonction de x donné et s'écrit $y = f(x)$. Cette écriture permet d'introduire la définition formelle suivante d'une fonction:

$$f = \{(x, y) \in X \times Y : y = f(x)\}. \tag{1-73}$$

De cette façon, le symbole f s'emploie pour définir une fonction et revêt deux significations:

1) f est l'ensemble qui a pour éléments les couples (x, y) intéressés par la correspondance;

2) $f(x)$ désigne $y \in Y$ correspondant à $x \in X$ donné.

La définition formelle de la fonction sous la forme de la relation (1-73) permet d'établir les procédés de définition de la fonction.

1. L'énumération de tous les couples (x, y) faisant partie de l'ensemble f, comme dans l'exemple 1-20. Ce mode de définition de la fonction peut être utilisé lorsque X est un ensemble fini. Pour plus de clarté, on dispose les couples (x, y) sous la forme d'un tableau.

2. Dans un grand nombre de cas, X et Y sont des ensembles de nombres réels ou complexes; alors $f(x)$ représente très souvent une

formule, c.-à-d. une expression réunissant quelques opérations mathématiques (addition, soustraction, division, prise de logarithme, etc.) qui doivent être effectuées sur $x \in X$ pour obtenir y.

Exemple 1-21. Soient $X = Y = R$ et $f = \{(x, y) \in R^2 : y = x^2\}$. Alors $f(x) = x^2$.

Quelquefois, les différents sous-ensembles de l'ensemble X de la fonction s'expriment au moyen de formules différentes. Supposons que $A_1, \ldots, A_n$ sont les sous-ensembles disjoints deux à deux de l'ensemble X. Soit $f_i(x)$ $(i = 1, \ldots, n)$ la formule définissant y pour $x \in A_i$. Alors la fonction $f(x)$ sera définie par l'expression

$$f(x) = \begin{cases} f_1(x) \text{ pour } x \in A_1; \\ \cdots\cdots\cdots \\ f_n(x) \text{ pour } x \in A_n. \end{cases} \tag{1-74}$$

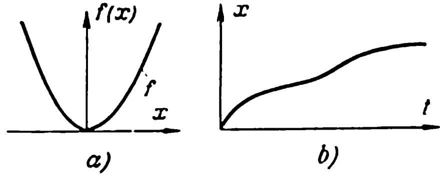


Fig. 1-14. Graphiques de fonctions

Ainsi, la fonction $y = f(x) = |x|$ peut être donnée sous la forme

$$y = \begin{cases} x \text{ pour } x \geqslant 0; \\ -x \text{ pour } x < 0. \end{cases}$$

3. Si X et Y sont les ensembles des nombres réels, les éléments $(x, y) \in f$ peuvent être représentés par des points sur le plan R^2. La totalité de ces points formera le *graphique* de la fonction $f(x)$. La figure 1-14,*a* donne le graphique de la fonction de l'exemple 1-21.

Dans les problèmes de cybernétique, on a souvent affaire à des fonctions du temps. Ces fonctions définissent l'application d'un nombre fini ou infini de points d'un certain intervalle de temps T sur l'ensemble des nombres réels $X \subseteq R$, ce qui s'écrit sous la forme

$$f : T \to X. \tag{1-75}$$

Désignant par t les éléments de l'ensemble T et par x ceux de l'ensemble X, on obtient la fonction $x = f(t)$ qui définit la variation de x avec le temps, comme il est montré, par exemple, sur la figure 1-14,*b*. Pour simplifier l'écriture, nous désignerons x en fonction du temps simplement par $x(t)$.

Si dans (1-72) $X = U \times V$, on aboutit à une fonction de deux variables u et v désignée par $f(u, v)$, où $u \in U$ et $v \in V$. La définition formelle de la fonction de deux variables réelles est:

$$f = \{(u, v, y) \in U \times V \times Y : y = f(u, v)\}. \tag{1-76}$$

On définit de la même façon les fonctions de trois variables et d'un nombre plus grand de variables.

Etant un cas particulier de la correspondance, la fonction vérifie les notions de fonction inverse et de composée de fonctions qui ont été établies pour la correspondance. Si f et g sont deux fonctions sur l'ensemble R^2 et que

$$f : X \rightarrow Y, \quad g : Y \rightarrow Z, \tag{1-77}$$

les fonctions inverses seront:

$$f^{-1} : Y \rightarrow X, \quad g^{-1} : Z \rightarrow Y. \tag{1-78}$$

La composée des fonctions f et g

$$f \circ g : X \rightarrow Z \tag{1-79}$$

pour tout $x \in Z$ définit $z \in Z$ qui s'écrit

$$z = (f \circ g)\, x = g\,[f(x)]. \tag{1-80}$$

La fonctionnelle est une notion plus générale que la fonction. La *fonctionnelle* établit une dépendance entre un ensemble de nombres d'une part et un certain ensemble de fonctions d'autre part. En d'autres termes, la fonctionnelle définit la relation entre un nombre et une fonction. Pour donner un exemple de la fonctionnelle, citons une intégrale définie de la forme

$$J(f) = \int_a^b f(x)\, dx.$$

Nous voyons que la fonctionnelle $J(f)$ est un nombre dépendant d'une fonction $f(x)$, celle-ci étant choisie dans un ensemble donné de fonctions.

L'opérateur est une notion encore plus générale. L'*opérateur* établit la correspondance entre deux ensembles de fonctions, de sorte qu'à chaque fonction d'un ensemble correspond une certaine fonction de l'autre. Désignant par p l'opérateur de dérivation, on écrit la relation entre la dérivée $f'(x) = df(x)/dx$ et la fonction $f(x)$ sous la forme d'un opérateur

$$f'(x) = p\,[f(x)].$$

1-6. RELATIONS

a) Propriétés des relations

Comme on l'a déjà indiqué, le terme « relation » s'emploie pour désigner certains types d'applications définies sur un seul et même ensemble. L'emploi de ce terme amène l'introduction d'une notation spéciale.

Supposons que l'application (X, Γ) soit une relation. Considérons l'élément $y \in \Gamma x$. Le fait que l'élément y est lié par la relation Γ à l'élément x s'écrira

$$y\Gamma x. \tag{1-81}$$

C'est ainsi que dans l'exemple 1-18 le symbole Γ veut dire « être enfants de l'homme donné ».

N o t a. Utilisant pour une application définie sur un seul et même ensemble la relation (1-64), on se rend compte que la relation est un couple d'ensembles (X, Γ) dans lequel $\Gamma \subseteq X^2$. Puisque l'ensemble X^2 a pour éléments les couples ordonnés, on peut dire que la relation est un ensemble de couples ordonnés. Chaque couple ne réunissant que deux éléments de l'ensemble X^2, on dit quelquefois que la relation est *binaire*.

Il est possible d'introduire une notion plus générale de relation en désignant sous ce terme un couple d'ensembles (X, Γ) où $\Gamma \subseteq X^n$. L'ensemble X^n a pour élements des n-uplets ordonnés, ce qui permet de le nommer relation *n-aire*. En particulier, un ensemble de triplets ordonnés sera une relation *ternaire*. Dans la suite, en disant « relation », l'on sous-entendra de façon tacite une relation *binaire*.

Il existe plusieurs catégories de relations en fonction de certaines propriétés que ces relations possèdent ou ne possèdent pas.

On considère ci-dessous les six propriétés fondamentales des relations en admettant que x, y et z sont des éléments quelconques d'un ensemble X.

Réflexivité : $x\Gamma x$ est vrai ; antiréflexivité : $x\Gamma x$ est faux ; symétrie : $x\Gamma y \rightarrow y\Gamma x$; antisymétrie : $x\Gamma y$ et $y\Gamma x \rightarrow x = y$; asymétrie : si $x\Gamma y$ est vrai, $y\Gamma x$ est faux ; transitivité : $x\Gamma y$ et $y\Gamma z \rightarrow x\Gamma z$.

Considérons, sur la base de ces propriétés, quelques catégories importantes de relations.

b) Relation d'équivalence

Certains éléments d'un ensemble peuvent être considérés comme équivalents si, dans une considération donnée, il est possible de substituer à n'importe lequel de ces éléments un autre élément quelconque. On dit alors que les éléments en question sont liés par la *relation d'équivalence*.

Voici quelques exemples de relations d'équivalence :

relation « faire partie d'une même année d'études (promotion) » définie sur l'ensemble des étudiants de la faculté;

relation « avoir le même reste après division par 3 » sur l'ensemble des entiers naturels ;

relation de parallélisme sur l'ensemble des droites du plan ;

relation de similitude sur l'ensemble des triangles, etc.

Pour donner une définition bien nette à la relation d'équivalence,

admettons que le terme de « relation d'équivalence » ne se justifie que si les trois conditions ci-dessous sont remplies:

1) chaque élément est équivalent à lui-même;

2) la proposition « deux éléments sont équivalents » reste vraie même si l'on ne précise pas lequel des éléments est le premier et lequel est le deuxième;

3) deux éléments équivalents à un troisième sont équivalents entre eux.

Adoptons, pour désigner l'équivalence, le symbole $\equiv$. La définition générale de l'équivalence s'écrit alors, au moyen des trois conditions énumérées ci-dessus, sous la forme de trois expressions:

1) $x \equiv x$ (réflexivité);

2) $x \equiv y \rightarrow y \equiv x$ (symétrie);

3) $x \equiv y$ et $y \equiv z \rightarrow x \equiv z$ (transitivité).

Ainsi donc, pour que la relation Γ soit une relation d'équivalence, il faut qu'elle soit réflexive, symétrique et transitive.

La relation d'équivalence se trouve intimement liée à la partition d'un ensemble (§ 1-2). Soit X l'ensemble sur lequel est définie une relation d'équivalence. Par exemple, X est l'ensemble des étudiants de la promotion, tandis que la relation d'équivalence s'exprime par la relation « être membres d'un même groupe d'études ». Le sous-ensemble des éléments équivalents à un certain élément $x \in X$ sera dit *classe d'équivalence*. De cette façon, le groupe d'études dans lequel fait ses études l'étudiant Ivanov sera la classe d'équivalence équivalente à l'étudiant Ivanov.

Soit J un ensemble d'indices. Désignons par $\{A_j \subseteq X : j \in J\}$ l'ensemble des classes d'équivalence pour l'ensemble X. Il est évident que tous les éléments d'une classe d'équivalence sont équivalents entre eux (en vertu de la propriété de transitivité) et un élément quelconque $x \in X$ ne peut se trouver que dans une classe et une seule. Or, X représente alors la réunion des ensembles disjoints A_j, de sorte que le système complet des classes $\{A_j \subseteq X : j \in J\}$ est une partition de X. Ainsi donc, à chaque relation d'équivalence sur l'ensemble X correspond une certaine partition de X en classes A_j.

On dit que la relation d'équivalence sur X et la partition de X en classes sont *conjuguées* lorsque, pour tout x et y sur X, la relation $x \equiv y$ se vérifie si et seulement si x et y appartiennent à une seule et même classe A_j de cette partition. Comparant les exemples du paragraphe donné à ceux du paragraphe 1-2, on arrive à se faire une idée plus nette de la liaison existant entre la relation d'équivalence et la partition d'un ensemble.

En qualité de symbole général de la relation d'équivalence, on emploie le symbole $\equiv$ (quelquefois $\sim$). Par contre, pour désigner des relations particulières d'équivalence, on fait usage d'autres symboles, notamment: $=$ pour désigner une égalité, $\parallel$ pour un parallélisme, $\rightleftarrows$ pour une équivalence logique.

c) Relation d'ordre

On a très souvent affaire aux relations qui définissent l'ordre successif des éléments d'un ensemble. C'est ainsi que nous distinguons les notions « avant » et « après » quand les éléments de l'ensemble sont les états d'un système dynamique, et les notions « plus grand que » (symbole $>$) et «plus petit que» (symbole $<$) quand ces éléments sont des nombres. Les notions d'ensemble et de sous-ensemble se distinguent à l'aide des symboles $\subseteq$ ou $\subset$.

Dans tous ces cas il est possible de disposer les éléments de l'ensemble X ou des groupes d'éléments dans un *ordre* déterminé, ou, en d'autres mots, d'introduire une relation d'ordre dans l'ensemble X.

On distingue la *relation d'ordre* (ou *d'ordre partiel*) désignée par le symbole $\leqslant$ (cas particuliers $\leqslant$, $\subseteq$) et la *relation de bon ordre* (ou *d'ordre stricte*) désignée par le symbole $<$ (cas particuliers $<, \subset, \rightarrow$). Définissons ces relations en énumérant les propriétés qu'elles doivent posséder.

On parle d'une relation d'ordre partiel si elle possède les trois propriétés suivantes:
$x \leqslant x$ est vrai (réflexivité);
$x \leqslant y$ et $y \leqslant x \rightarrow x = y$ (antisymétrie);
$x \leqslant y$ et $y \leqslant z \rightarrow x \leqslant z$ (transitivité).

On parle d'une relation de bon ordre si elle possède les trois propriétés suivantes:
$x < x$ est faux (antiréflexivité);
$x < y$ et $y < x$ s'excluent (asymétrie);
$x < y$ et $y < z \rightarrow x < z$ (transitivité).

On dit que l'ensemble X est *ordonné* si deux éléments quelconques x et y de cet ensemble sont comparables, c.-à-d. si l'on a

$$x < y \text{ ou } x = y \text{ ou } y < x.$$

d) Relation de dominance

Dans les cas où X désigne un ensemble d'hommes ou de groupes d'hommes, on rencontre quelquefois une relation dite *de dominance*. On dit que x domine y et l'on écrit $x \gg y$ si x s'avère supérieur à y sous un rapport quelconque. Par exemple, x peut être le sportif ou l'équipe sportive qui a gagné le match au sportif ou à l'équipe sportive y, ou bien une personne qui a de l'autorité auprès de la personne y, ou encore une propriété que nous préférons à la propriété y.

Nous dirons que les éléments de X sont liés par la relation de dominance si ces éléments possèdent les deux propriétés suivantes:

1) aucun individu ne peut se dominer soi-même, c.-à-d. que $x \gg x$ est faux (antiréflexivité);

2) dans chaque couple d'individus un individu domine absolument l'autre, c.-à-d. que $x \gg y$ et $y \gg x$ s'excluent réciproquement (asymétrie).

Lorsqu'il s'agit de la relation de dominance, la propriété de transitivité n'intervient pas. En effet, si l'équipe x a gagné le match à l'équipe y et cette équipe y a gagné le match à l'équipe z, cela ne veut point dire que l'équipe x gagnera à coup sûr dans le match avec l'équipe z.

1-7. QUELQUES NOTIONS D'ALGÈBRE SUPÉRIEURE

a) Groupes, anneaux, corps

Avec l'évolution des mathématiques, les opérations algébriques (addition, multiplication, division) applicables initialement aux nombres rationnels furent étendues à toute une série d'autres objets: nombres complexes, vecteurs, matrices, etc. Les règles d'exécution de ces opérations varient selon l'objet. Toutefois ces opérations possèdent des propriétes communes dont la connaissance permet d'établir s'il est possible d'effectuer l'opération donnée sur une classe concrète d'objets. L'établissement de ces propriétés permet de dégager les notions d'opération algébrique, groupe, anneau, corps.

Soit X un ensemble quelconque. Faisant correspondre à tout couple ordonné $(a, b) \in X^2$, de façon univoque, un certain élément c appartenant au même ensemble X, on dit qu'une *opération algébrique* est définie sur X. L'opération ainsi définie s'appelle *multiplication* ou *addition* et s'écrit

$$c = ab \quad \text{ou} \quad c = a + b. \tag{1-82}$$

On dit que l'opération algébrique est *associative* si pour tout $a, b, c \in X$ on a

$$(ab)\,c = a\,(bc) \quad \text{ou} \quad (a + b) + c = a + (b + c).$$

L'ensemble X avec une opération associative définie sur cet ensemble (désignée en attendant sous le terme de multiplication) s'appelle *demi-groupe*. Un demi-groupe est dit *groupe* si:

1) il existe dans X un élément e tel que, pour tout $a \in X$,

$$ae = ea = a; \tag{1-83}$$

2) il existe pour tout $a \in X$ un élément a^{-1} tel que

$$aa^{-1} = a^{-1}a = e. \tag{1-84}$$

L'élément e porte le nom d'*unité* du groupe; l'élément a^{-1} est l'*inverse* de a. Si l'opération définie sur le groupe est l'addition, on dit que e est le *zéro* du groupe (symbole 0) et l'élément a^{-1} l'*opposé* à a (symbole $-a$). Les éléments 0 et $-a$ vérifient les relations

$$a + 0 = 0 + a = a, \; a + (-a) = 0. \tag{1-85}$$

Si X est un ensemble fini, le groupe s'appelle *fini*. Voici quelques exemples de groupes:

ensemble d'entiers vis-à-vis de l'opération d'addition;

ensemble de nombres pairs vis-à-vis de l'opération d'addition;

ensemble des nombres rationnels non nuls vis-à-vis de l'opération de multiplication;

ensemble des vecteurs du plan vis-à-vis de l'opération d'addition des vecteurs.

Cependant, les nombres entiers ne constituent pas de groupe vis-à-vis de l'opération de multiplication, car un nombre entier qui diffère de ± 1 n'a pas d'inverse entier. Pour donner un exemple de groupe fini, citons l'ensemble $\{1, -1, i, -i\}$ dans lequel $i = \sqrt{-1}$ vis-à-vis de l'opération de multiplication.

Supposons que deux opérations algébriques, l'addition et la multiplication, soient définies simultanément sur un ensemble X. Supposons en outre que l'opération d'addition soit commutative ($a + b = b + a$) et que l'opération de multiplication soit liée à celle d'addition par des lois distributives

$$a(b + c) = ab + ac \quad \text{et} \quad (b+c)a = ba+ca. \qquad (1\text{-}86)$$

On dit alors que l'ensemble en question est un *anneau*. Un anneau peut ne pas avoir d'unité ni d'éléments inverses. Si l'anneau possède bien une unité, c'est un *anneau unitaire*. Parmi les anneaux, on compte les ensembles des nombres entiers, des nombres rationnels, des nombres réels, des nombres complexes vis-à-vis des opérations classiques d'addition et de multiplication.

Un anneau dans lequel pour tout $a \neq 0$ et pour tout b il y a exactement un élément x tel que $ax = b$ s'appelle *corps*. L'élément x s'appelle *quotient* de b par a et se désigne par $x = b/a$. Pour donner des exemples de corps, citons l'anneau des nombres rationnels, l'anneau des nombres réels, l'anneau des nombres complexes.

b) Isomorphisme. Homomorphisme. Modèles

Dans les études scientifiques et pratiques, un rôle remarquable revient à la simulation des systèmes et situations réels. On entend sous ce terme l'établissement d'une relation d'équivalence entre deux systèmes, dont chacun peut être ou bien une réalité objective, ou bien une abstraction. S'il s'avère que le premier système se prête mieux à l'étude que le deuxième, on se fait une idée des propriétés du deuxième en observant le comportement du premier. Dans ce cas le système choisi pour l'étude porte le nom de *modèle*.

Un modèle est dit *isomorphe* (de même forme) si le système réel et le modèle coïncident d'une manière très complète, élément par

élément. C'est le cas d'un négatif photographique et de l'épreuve tirée de ce négatif, d'un dessin et de la pièce exécutée d'après ce dessin, des processus quelconques se produisant dans un système réel et de la solution de l'équation décrivant le comportement du système.

Or, les modèles isomorphes s'avèrent très souvent excessivement compliqués et peu commodes pour l'utilisation pratique. Il y a plus d'intérêt à créer des modèles permettant de définir uniquement les aspects essentiels du comportement des systèmes réels, sans entrer inutilement dans le détail. C'est le cas, par exemple, d'une carte géographique vis-à-vis du terrain qu'elle représente.

Les modèles dont les différents éléments ne correspondent qu'à des parties plus ou moins grosses du système réel, sans qu'il y ait correspondance parfaite élément par élément entre modèle et système, sont dits *modèles homomorphes*.

L'isomorphisme et l'homomorphisme sont susceptibles d'une définition mathématique rigoureuse aux termes de la théorie des groupes.

Soient X et Y des groupes. Lorsque les éléments de ces groupes sont liés par une correspondance biunivoque telle que, pour n'importe quels éléments a, $b \in X$ et les éléments a', $b' \in Y$ qui leur correspondent, à l'élément $c = ab$ correspondra l'élément c' égal à $a'b'$, on dira que les groupes sont isomorphes.

Les groupes isomorphes ne peuvent différer que par la nature de leurs éléments et peut-être par le genre des opérations définies sur le groupe; cependant toutes les propriétés des groupes isomorphes entre eux, qui résultent des propriétés des opérations définies sur ces groupes et qui ne dépendent pas de la nature des éléments du groupe, sont communes.

Dans les groupes homomorphes, la correspondance entre les groupes est unilatérale. On dit que l'application du groupe X dans le groupe Y est *homomorphe* si à chaque élément de X correspond de façon univoque un élément déterminé de Y, de sorte que si aux éléments a, $b \in X$ correspondent les éléments a', $b' \in Y$, à l'élément $ab = c$ corresponde l'élément $c' = a'b'$. Dans le cas général, étant donné un homomorphisme de la forme $X \to Y$, il est possible que différents éléments du groupe X passent à l'élément donné du groupe Y; il est tout aussi possible qu'aucun élément n'y passe.

PROBLÈMES AU CHAPITRE PREMIER

1-1. Quel est l'ensemble $Y \setminus X$ de l'exemple 1-1?

1-2. Désigner par hachure l'ensemble $Y \setminus X$ de l'exemple 1-3.

1-3. Soient R l'ensemble des nombres réels et

$$X = \{x \in R : 0 \leqslant x \leqslant 1\}, \quad Y = \{y \in R : 0 \leqslant y \leqslant 2\}.$$

Quels sont les ensembles $X \cup Y$, $X \cap Y$, $X \setminus Y$?

1-4. Tracer les figures représentatives des ensembles $A = \{(x, y) \in R^2 : x^2 + y^2 \leqslant 1\}$ et $B = \{(x, y) \in R^2 : x^2 + (y - 1)^2 \leqslant 1\}$.

Par quelles figures sont représentés les ensembles $A \cup B$, $A \cap B$, $R^2 \setminus A$?

1-5. Utilisant (1-34), démontrer les identités $X \cap \varnothing = \varnothing$; $X \cup \varnothing = X$; $I \cap X = X$; $I \cup X = I$.

1-6. Représenter sur le plan réel $R^2 = R \times R$ les ensembles $X \times Y$ et $Y \times X$ du problème 1-3.

1-7. Représenter géométriquement les ensembles $A \times R$ et $R \times A$, où $A = [2, 3]$.

1-8. Etant donné l'ensemble $M = \{(x, y) \in R^2 : (x-2)^2 + y^2 = 1\}$, chercher $\mathrm{pr}_1\, M$ et $\mathrm{pr}_2\, M$.

1-9. Soient $I = \{x_1, x_2, x_3\}$ l'ensemble universel, et $X = \{x_1, x_2\}$, $Y = \{x_2, x_3\}$, $Z = \{x_3\}$ ses sous-ensembles. Définir par énumération les ensembles suivants:

$$X \times X;\quad Z \times Z;\quad X \times Y;\quad Y \times X;\quad X \times Y \cap Y \times X;\quad X \times Y \cup Y \times X.$$

1-10. Soient X, Y, Z les sous-ensembles de l'ensemble R^2 égaux à $X = \{(x, y) : x \geqslant 0\}$; $Y = \{(x, y) : y \geqslant 0\}$; $Z = \{(x, y) : x + y \geqslant 1\}$. Donner la représentation géométrique des ensembles

$$X;\ Y;\ Z;\ X \cup Y;\ \overline{X \cup Y};\ X \cap Y;\ \overline{X \cap Y};\ X \cap Z;\ \overline{X \cap Z};$$
$$X \cap Y \cap Z;\quad X \cap Y \cap \overline{Z}.$$

1-11. Quel est l'ensemble $X \times Y$ si X et Y sont les sous-ensembles de l'ensemble R^2 égaux à $X = \{(x, y) : 2x + y = 1\}$; $Y = \{(x, y) : x - y = 0\}$?

1-12. Ecrire pour l'exemple 1-13 toutes les seize correspondances. Définir pour chacune d'entre elles $\mathrm{pr}_1 Q$ et $\mathrm{pr}_2 Q$.

1-13. Trouver pour la fonction f de l'exemple 1-21 la fonction inverse f^{-1}.

1-14. Soient f et g les fonctions sur un ensemble R^2 telles que $f = \{(x, y) : y : x^2\}$ et $g = \{(y, z) : z = \sin y\}$. Trouver la composée de ces fonctions $f \circ g$.

CHAPITRE 2

ÉLÉMENTS DE LA THÉORIE DES GRAPHES

2-1. DÉFINITIONS PRINCIPALES DE LA THÉORIE DES GRAPHES

a) Définition d'un graphe en termes de la théorie des ensembles

La meilleure façon de se faire une idée de ce qu'est un *graphe* consiste à imaginer un ensemble de points d'un plan X, appelés *sommets*, et un ensemble de tronçons orientés U reliant tous les sommets ou certains d'entre eux et appelés *arcs* [12 à 14]. Du point de vue mathématique, un graphe G est un couple d'ensembles X et U:

$$G = (X, U). \qquad (2\text{-}1)$$

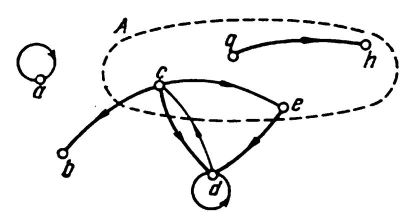


Fig. 2-1. Vue générale d'un graphe

On voit à la figure 2-1 un graphe dont les sommets sont les points a, b, c, d, e, g, h, et les arcs, les tronçons (a, a), (c, b), (c, d), (d, c), (d, d), (c, e), (e, d), (g, h). Voici quelques exemples de graphes: relations de paternité et de maternité dans un ensemble des hommes (voir fig. 1-13), carte routière, schéma des connexions d'appareils électriques, relations de supériorité de certains participants d'un tournoi par rapport aux autres, et ainsi de suite.

Il est quelquefois plus commode de donner au graphe une autre définition. On peut admettre que l'ensemble des arcs dirigés U reliant les éléments d'un ensemble X est l'application de ce dernier ensemble dans lui-même. On considère donc que le graphe est défini si l'on connaît l'ensemble de ses sommets X et le mode d'application Γ de l'ensemble X dans X. Ainsi donc, un graphe G est un couple (X, Γ) réunissant un ensemble X et une application Γ définie sur cet ensemble:

$$G = (X, \Gamma). \qquad (2\text{-}2)$$

Par exemple, pour le graphe représenté sur la figure 2-1 l'application Γ trouve la définition suivante:

$$\Gamma a = a;\ \Gamma b = \varnothing;\ \Gamma c = \{b, d, e\};\ \Gamma d = \{d, c\};\ \Gamma e = d;$$
$$\Gamma g = h; \quad \Gamma h = \varnothing.$$

On se rend compte sans peine que cette définition du graphe coïncide entièrement avec celle d'une relation sur un ensemble.

Il y a parfois intérêt à représenter les graphes sous la forme de matrices, en particulier des matrices associées à un graphe et des matrices d'incidence. Donnons d'abord deux définitions.

Deux sommets x et y sont *adjacents* s'ils sont distincts et s'il existe un arc allant de x à y.

Un arc u prenant naissance ou se terminant en x est dit *incident* au sommet x.

Désignons par $x_1, \ldots, x_n$ les sommets d'un graphe, et par $u_1, \ldots, u_m$ ses arcs. Introduisons les nombres:

$$r_{ij} = \begin{cases} 1 \text{ s'il existe un arc reliant les sommets } i \text{ et } j; \\ 0 \text{ si cet arc est absent.} \end{cases}$$

Une matrice carrée $R = \| r_{ij} \|$ d'ordre $n \times n$ porte le nom de *matrice associée* à un graphe.

Introduisons ensuite les nombres

$$s_{ij} = \begin{cases} +1 \text{ si } u_j \text{ est incident à } x_i \text{ vers l'extérieur;} \\ -1 \text{ si } u_j \text{ est incident à } x_i \text{ vers l'intérieur;} \\ 0 \text{ si } u_j \text{ n'est pas incident à } x_i. \end{cases}$$

Une matrice $S = \| s_{ij} \|$ d'ordre $n \times m$ est dite *matrice d'incidence* pour les arcs du graphe.

Les matrices d'incidence telles qu'on vient de les définir ne sont applicables qu'à des graphes sans boucles. Si le graphe comporte des boucles, il convient de partager la matrice en deux demi-matrices, positive et négative.

Introduisons quelques notions et définitions utiles pour la description de certaines catégories de graphes.

Le *sous-graphe* G_A d'un graphe $G = (X, \Gamma)$ est un graphe qui n'englobe que la partie des sommets du graphe G formant l'ensemble A avec les arcs réunissant ces sommets: c'est le cas du domaine délimité par la courbe en trait pointillé à la figure 2-1. La définition mathématique du sous-graphe G_A est

$$G_A = (A, \Gamma_A), \tag{2-3}$$

où

$$A \subseteq X, \quad \Gamma_A x = (\Gamma x) \cap A. \tag{2-4}$$

Un *graphe partiel* G_Δ par rapport à un graphe $G = (X, \Gamma)$ est un graphe ne contenant qu'une partie des arcs du graphe G, c.-à-d. défini par la condition

$$G_\Delta = (X, \Delta), \tag{2-5}$$

où

$$\Delta x \subseteq \Gamma x. \tag{2-6}$$

A la figure 2-1 le graphe formé par les arcs tracés en trait gras est un graphe partiel.

Exemple 2-1. Soit $G = (X, \Gamma)$ la carte routière de l'Union Soviétique. Alors la carte routière de la région de Tambov est un sous-graphe, et la carte routière sommaire de l'Union Soviétique ne montrant que les artères principales est un graphe partiel.

Deux autres notions importantes sont celles du chemin et du circuit. On a défini un arc comme un tronçon dirigé reliant deux sommets. L'arc reliant les sommets a et b et dirigé de a vers b se note $u = (a, b)$.

Le *chemin* dans un graphe G est une succession d'arcs $\mu = (u_1, \ldots, u_k)$ où l'origine de l'un est l'extrémité du précédent. Le chemin μ qui passe par les sommets successifs $a, b, \ldots, m$ est désigné par $\mu = (a, b, \ldots, m)$. La *longueur du chemin* $\mu = (u_1, \ldots, u_k)$ est la quantité $l(\mu) = k$ égale au nombre d'arcs formant le chemin μ. Un chemin peut être fini ou infini. Pour le chemin infini on a $l(\mu) = \infty$. Un chemin n'empruntant jamais deux fois le même arc est dit *simple*. Un chemin ne passant jamais deux fois par le même sommet est dit *élémentaire*.

Le *circuit* est un chemin fini $\mu = (x_1, \ldots, x_k)$ dont le sommet initial x_1 se confond avec le sommet terminal x_k. Le circuit est dit *élémentaire* s'il a tous ses sommets distincts (à l'exception de l'origine et de l'extrémité qui coïncident). Un circuit de longueur unité formé par un arc de la forme (a, a) est appelé *boucle*. On voit sur la figure 2-1 un chemin (e, d, c, b), un circuit (c, e, d, c) et une boucle (d, d).

Il y a des cas où, considérant un graphe, on fait abstraction de l'orientation de ses arcs. On dit alors qu'il s'agit d'un graphe *non orienté*. Pour un graphe non orienté, les notions d'arc, de chemin et de circuit cèdent la place à celles d'arête, de chaîne et de cycle. L'*arête* est un tronçon reliant deux sommets. Le graphe de la figure 2-1 possède huit arcs mais sept arêtes. La *chaîne* est une succession d'arêtes. Le *cycle* est une chaîne finie qui a l'origine et l'extrémité en un même point.

A la notion de graphe non orienté est liée une propriété fort importante de *connectivité* du graphe. Un graphe est *connexe* si deux quelconques de ses sommets sont unis par une même chaîne. Un graphe non connexe G peut être partagé en sous-graphes G_i de telle façon que tous les sommets dans chaque sous-graphe soient connexes mais que les sommets des différents sous-graphes ne le soient pas. De tels sous-graphes G_i portent le nom de *composantes connexes* du graphe G.

Pour trouver si un graphe orienté est connexe ou non, il y a lieu de faire abstraction de l'orientation des arcs. Le graphe représenté sur la figure 2-1 n'est pas connexe. Or, son sous-graphe comprenant les sommets b, c, d, e est connexe. On dit parfois d'un graphe orienté qu'il est *fortement connexe*: on entend par là que, pour n'importe quels deux sommets x et y $(x \neq y)$ il y a un chemin allant de x à y.

Un cas particulier fort important d'un graphe non orienté est l'arbre. On entend par *arbre* un graphe non orienté fini connexe sans cycles; quelques exemples d'arbres sont représentés à la figure 2-2.

Etant donné un ensemble de sommets a, b, c, . . ., l'arbre se construit de la façon suivante. Choisissons un sommet a en qualité de point de départ : ce sera la *racine* de l'arbre. A partir de ce sommet, traçons des arêtes jusqu'aux sommets les plus proches b, c, d, . . ., et à partir de ceux-ci, jusqu'aux sommets voisins e, f, g, h, . . ., et ainsi de suite. Ainsi donc, un arbre se construit en ajoutant successivement des arêtes dans ses sommets. Ceci permet d'établir un rapport entre le nombre de sommets et celui d'arêtes d'un arbre.

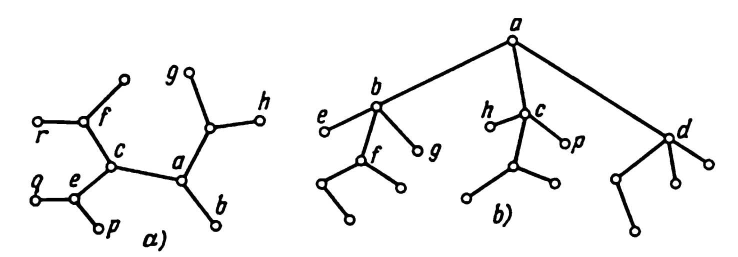


Fig. 2-2. Exemples d'arbres

L'arbre le plus élémentaire se compose de deux sommets liés entre eux par une arête. Chaque fois que nous y ajoutons une arête, nous ajoutons par là même encore un sommet qui se trouve au point terminal de cette arête. Donc, un arbre à n sommets possède $n - 1$ arêtes.

b) Relation d'ordre et relation d'équivalence sur un graphe

Nous venons de voir que le graphe permet de donner une représentation géométrique commode des relations sur un ensemble. Aussi la théorie des graphes et la théorie des relations sur un ensemble se complètent-elles mutuellement.

Nous admettrons qu'un graphe $G = (X, \Gamma)$ est muni d'une relation d'ordre si ses deux sommets quelconques x et y, tels que $x \leqslant y$, peuvent être liés par un chemin allant de x à y. On dira alors que le sommet x précède le sommet y, ou que le sommet y suit le sommet x.

Montrons que cette définition tient compte de toutes les propriétés de la relation d'ordre sur un graphe.

Réflexivité. La condition

$$x \leqslant x \text{ est vrai} \tag{2-7}$$

veut dire que le sommet est équivalent à lui-même, soit $x \equiv x$. D'autre part, il n'est pas moins légitime de considérer cette condition comme celle d'existence d'un chemin allant de x à x, c.-à-d. de la boucle au sommet x (fig. 2-3,*a*).

Transitivité. La condition

$$x \leqslant y, \quad y \leqslant z \rightarrow x \leqslant z \tag{2-8}$$

signifie que l'on rencontre successivement les sommets x, y, z en parcourant un seul et même chemin (fig. 2-3,*b*).

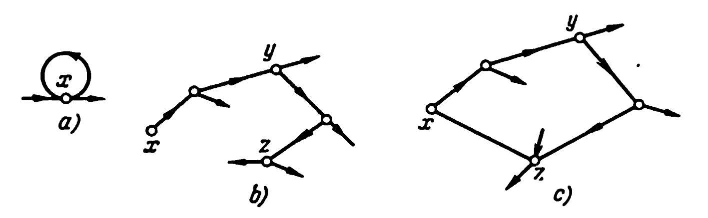


Fig. 2-3. Illustration des propriétés de la relation d'ordre

Antisymétrie. Montrons que la condition

$$x \leqslant y, \quad y \leqslant x \rightarrow x \equiv y \tag{2-9}$$

est vraie. On voit du premier membre de cette expression qu'il existe le chemin de x à y et aussi le chemin de y à x. Or, cela équivaut à dire que le graphe comprend un circuit passant par les sommets x et y (fig. 2-3,*c*).

Il ressort du second membre de (2-9) que les sommets situés sur un même circuit sont équivalents. Assimilons cette conclusion à une définition de l'équivalence sur un graphe et montrons que cette définition tient compte de toutes les trois conditions de la relation d'équivalence. La condition de réflexivité $x \equiv x$ et celle de symétrie $x \equiv y \rightarrow y \equiv x$ sont évidentes et découlent de la définition de l'équivalence donnée ci-dessus. La condition de transitivité $x \equiv y$, $y \equiv z \rightarrow x \equiv z$ est aussi évidente, car elle suppose qu'un graphe muni d'un circuit de sommets x et y et d'un circuit de sommets y et z comporte sûrement un circuit de sommets x et z (voir fig. 2-3,*c*).

De cette façon, la relation d'ordre et la relation d'équivalence définissent conjointement un certain graphe.

Un graphe peut aussi être muni d'une relation de bon ordre. Dans ce cas, pour tout couple de sommets x et y vérifiant la condition $x < y$, il existe un chemin de x à y. De même que dans le cas précédent, la condition de transitivité $x < y$, $y < z \rightarrow x < z$ veut dire que x, y et z sont les sommets successifs d'un seul et même chemin. La condition d'antiréflexivité ($x < x$ est faux) démunit le

graphe de boucles, et la condition d'asymétrie ($x < y$, $y < x$ s'excluant réciproquement) le démunit de circuits.

On voit donc que la relation de bon ordre définit un graphe sans circuits.

c) Caractéristiques des graphes

La résolution de nombreux problèmes techniques par les méthodes de la théorie des graphes se ramène à définir telles ou telles caractéristiques des graphes. Bien que le présent ouvrage ne puisse englober les applications techniques de la théorie des graphes, la connaissance des principales caractéristiques des graphes peut s'avérer d'une grande utilité lors de l'étude des matières les plus diverses.

Nombre cyclomatique. Soit G un graphe non orienté ayant n sommets, m arêtes et r composantes connexes. Le nombre cyclomatique de ce graphe G est le nombre

$$\nu(G) = m - n + r.$$

Ce nombre a la signification physique remarquable: il est égal au plus grand nombre de cycles indépendants que le graphe puisse avoir. Lors du calcul des réseaux électriques, on se sert du nombre cyclomatique pour déterminer le nombre de circuits indépendants.

Nombre chromatique. Soit p un entier naturel. On dit d'un graphe G qu'il est p-chromatique s'il est possible de peindre ses sommets en p couleurs différentes de telle façon qu'aucun de ses sommets ne soit de la même couleur qu'un sommet adjacent. Le plus petit nombre p pour lequel le graphe est p-chromatique est appelé nombre chromatique du graphe et est noté $\gamma(G)$.

Si $\gamma(G) = 2$, le graphe est dit bichromatique. Pour qu'un graphe soit bichromatique, il faut et il suffit qu'il soit exempt de cycles de longueur impaire.

Le nombre chromatique joue un rôle très important lorsqu'il s'agit d'utiliser de façon optimale les cellules de la mémoire en établissant le programme. Il est à noter cependant que sa détermination, sauf s'il s'agit d'un graphe bichromatique, est difficile et exige assez souvent la mise en œuvre d'une machine à calculer électronique.

Ensemble stable intérieurement. Un ensemble $S \subseteq X$ sur un graphe $G = (X, \Gamma)$ est dit stable intérieurement s'il n'y a pas dans S deux sommets adjacents, c.-à-d. si, pour tout $x \in S$, on a $\Gamma x \cap S = \varnothing$.

Un ensemble stable intérieurement contenant le plus grand nombre d'éléments porte le nom du plus grand ensemble stable intérieurement, et le nombre d'éléments de cet ensemble s'appelle nombre de stabilité interne du graphe G. Le plus grand ensemble stable intérieurement joue un rôle important dans la théorie des communications.

Ensemble stable extérieurement. Un ensemble $T \subset X$ sur un graphe $G = (X, \Gamma)$ est dit stable extérieurement si n'importe quel sommet extérieur à T est relié par des arcs à des sommets de T, c.-à-d. si pour tout $x \notin T$ on a $\Gamma x \cap T \neq \varnothing$.

Un ensemble stable extérieurement contenant le plus petit nombre d'éléments porte le nom du plus petit ensemble stable extérieurement, et le nombre d'éléments de cet ensemble s'appelle nombre de stabilité externe du graphe G.

2-2. PROBLÈME DU PLUS COURT CHEMIN

a) Position du problème

Dans des applications pratiques, on attache une importance considérable à la recherche du plus court chemin entre deux sommets d'un graphe non orienté connexe. C'est à un problème de ce genre qu'on peut ramener de nombreux cas où il s'agit de fixer l'itinéraire optimal (du point de vue de la distance, du temps, du coût) sur la carte de routes donnée, de choisir le procédé le plus économique pour faire passer un système dynamique d'un état à un autre, etc. Les mathématiques fournissent nombre de méthodes convenant à ces problèmes. Or, ce sont encore les méthodes relevant de la théorie des graphes qui s'avèrent bien souvent les moins encombrantes.

Dans sa forme générale, le problème du plus court chemin sur un graphe s'énonce comme suit. Soit un graphe non orienté $G = (X, U)$. Chaque arête de ce graphe se caractérise par un nombre $l(u) \geqslant 0$ dit longueur de l'arête. Suivant le cas, le nombre $l(u)$ peut se présenter comme la distance entre les sommets réunis par l'arête u, comme la durée ou le prix du voyage le long de cette arête, etc. Toute chaîne μ aura alors pour longueur

$$l(\mu) = \sum_{u \in \mu} l(u). \tag{2-10}$$

Il s'agit de chercher, pour deux sommets arbitraires a et b du graphe G, un chemin μ_{ab} tel que sa longueur totale soit aussi petite que possible.

Avant d'aborder la méthode générale de résolution de ce problème, arrêtons-nous sur la règle donnant solution du problème particulier dans lequel la longueur de chaque arête est égale à l'unité.

b) Recherche du plus court chemin dans un graphe à arêtes de longueur unité

On rencontre parfois des graphes dont les arêtes ont la même longueur prise pour l'unité. Les sommets d'un tel graphe représentent généralement les états d'un certain système dans lequel

tous les passages effectués en un seul pas sont, d'un certain point de vue, équivalents. Nous proposons au lecteur un problème qui se ramène au cas d'un graphe à arêtes de longueur unité. Ce problème illustre les modes de construction des graphes pour différents cas concrets.

Exemple 2-2. Problème de la tour de Hanoï. Dans une plaque sont plantées trois fiches. Sur la première fiche sont enfilées m rondelles dont le diamètre décroît de bas en haut. On demande de transférer les rondelles une à une à la troisième fiche en les disposant finalement dans l'ordre initial, de façon qu'à aucun

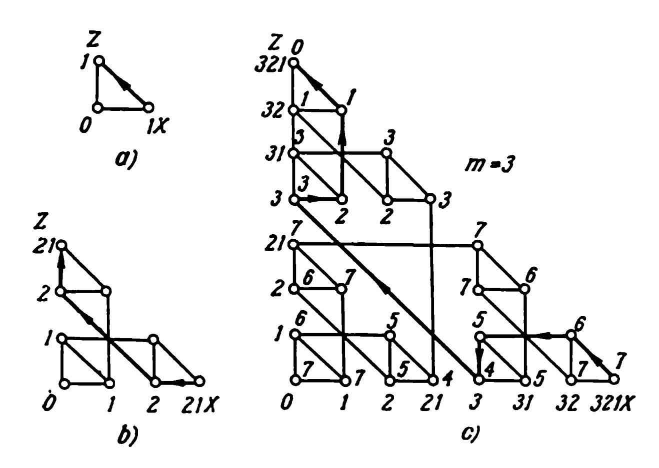


Fig. 2-4. Graphes des transferts dans le problème de la tour de Hanoï

moment et sur aucune fiche une rondelle de diamètre plus grand ne se trouve superposée à une rondelle de diamètre plus petit; la deuxième fiche sert de fiche intermédiaire. Une condition supplémentaire peut être imposée: la solution doit être obtenue avec le minimum de pas.

Affectons les rondelles de numéros dans l'ordre de décroissance de diamètre: m, $m-1$, . . ., 1. Désignons par X, Y et Z les ensembles des rondelles se trouvant à un pas quelconque respectivement sur la première, sur la deuxième et sur la troisième fiche. Ceci posé, il suffit de préciser les ensembles X et Z, l'ensemble Y étant complémentaire de X et Z dans l'ensemble de toutes les rondelles. Chacun des ensembles X ou Z peut être représenté par l'une des combinaisons suivantes de rondelles: 0, 1, 2, 21, 3, 31, 32, 321, 4, 41, 42, 421, 43, 431, 432, 4321, . . . On peut représenter ces combinaisons sous la forme de points sur les axes des X et Z, ainsi qu'il est montré sur la figure 2-4, de sorte que toute disposition des rondelles soit représentée par un certain point sur le plan (X, Z). Reliant ces points par des lignes qui montrent les transferts de rondelles possibles à chaque pas, on se trouve en présence d'un graphe non orienté permettant de trouver un chemin (et aussi le plus court chemin) de passage du point initial du graphe à son point terminal.

On peut construire le graphe en passant de m rondelles à $m+1$ rondelles. Pour $m=1$ les états possibles seront représentés par un ensemble $\{(1,0), (0,0), (0,1)\}$ auquel correspond le graphe de la figure 2-4,*a*; on y distingue le passage le plus court de l'état initial (1,0) à l'état final (0,1).

Pour obtenir la règle générale, admettons que l'on a déjà construit le graphe pour le cas de m rondelles; appelons-le m-graphe. Il est facile de voir qu'un tel graphe a la forme d'un triangle (fig. 2-5,*a*), sans qu'on connaisse toutefois, à cette étape, ses liaisons internes. Que devient ce graphe si l'on prend $m + 1$ rondelles?

Remarquons que la rondelle numéro $m + 1$ ne peut occuper, à toutes les fiches, que la position la plus basse. Les autres m rondelles peuvent se déplacer de manière arbitraire suivant le m-graphe, sans que la rondelle numéro $m + 1$ soit dérangée. Par conséquent, le graphe pour le cas de $m + 1$ rondelles

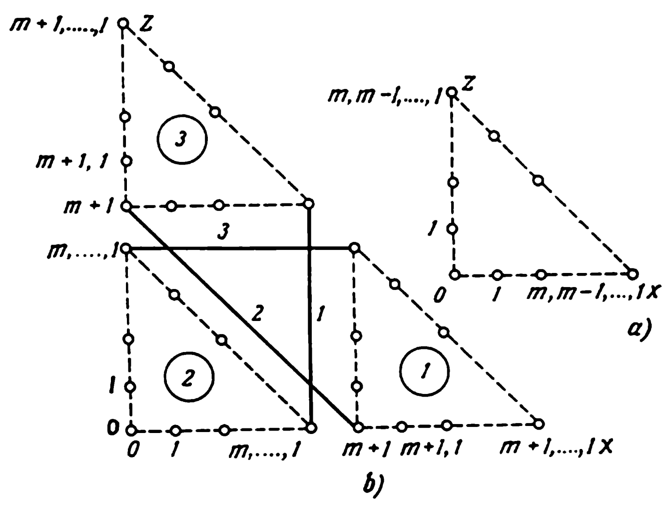


Fig. 2-5. Passage du m-graphe au $(m + 1)$-graphe

(fig. 2-5,*b*) se compose de trois m-graphes désignés par les chiffres cerclés *1*, *2* et *3* indiquant le numéro de la fiche sur laquelle est enfilée la rondelle numéro $m + 1$. Il reste de savoir quels sont les passages d'un m-graphe à l'autre qui correspondent au transfert de la rondelle numéro $m + 1$.

La rondelle numéro $m + 1$ ne peut être placée que sur une fiche libre; or, ce n'est possible que lorsque toutes les m rondelles de diamètre inférieur se trouvent réunies sur l'une des fiches. On voit sur la figure 2-5,*b* les transferts possibles de la rondelle numéro $m + 1$ désignés par les chiffres *1*, *2* et *3* qui indiquent le numéro de la fiche sur laquelle sont enfilées les m rondelles plus petites.

Le diagramme de la figure 2-5,*b* fournit le principe général de passage du m-graphe au $(m + 1)$-graphe. A la figure 2-4,*b* et *c*, sont résumés les graphes des transferts pour deux et trois rondelles.

Passons au problème de recherche dans le graphe du chemin le plus court entre le sommet initial et le sommet terminal. Les graphes examinés ci-dessus étant relativement simples, il serait facile de trouver le chemin le plus court par recensement des chemins possibles. Or, pour des graphes complexes, il importe de trouver une méthode générale.

La règle générale de recherche du chemin le plus court dans un graphe consiste en ce qu'on attribue à chaque sommet x_i un indice λ_i égal à la longueur du plus court chemin allant du sommet donné jusqu'au sommet terminal. Dans le cas où le graphe a toutes ses arêtes de longueur unité, les sommets sont marqués dans l'ordre suivant :

1) on attribue l'indice 0 au sommet terminal x_0 ;

2) tous les sommets liés par une arête au sommet terminal sont munis d'indice 1 ;

3) à tous les sommets non encore marqués reliés par une arête à un sommet d'indice λ_i on attribue l'indice $\lambda_i + 1$. Ce processus a lieu jusqu'à ce que le sommet initial soit marqué. A ce moment l'indice du sommet initial sera égal à la longueur du plus court chemin. Pour trouver le plus court chemin lui-même (l'itinéraire), il convient de partir du sommet initial dans la direction de décroissance des indices.

On voit à la figure 2-4,*c* un exemple de marquage des points et de recherche du plus court chemin pour $m = 3$.

Notons que le procédé de recherche du chemin de longueur minimale décrit ci-dessus est un cas particulier de la recherche de la solution optimale par la méthode de programmation dynamique. Aussi, après avoir étudié la programmation dynamique, serait-il profitable de revenir à l'exemple que l'on vient de considérer.

c) Recherche du plus court chemin dans un graphe à arêtes de longueur arbitraire

Le marquage des sommets d'un graphe avec des indices numériques devient plus compliqué si les arêtes du graphe sont de longueur arbitraire. Ceci tient à ce que, dans un graphe complexe, le chemin passant par le plus petit nombre de sommets s'avère quelquefois plus long que certains chemins de détour. C'est ainsi que, dans le graphe de la figure 2-6, le chemin direct du sommet marqué avec un astérisque au sommet terminal a la longueur $l = 12$, tandis que le chemin de détour passant par le sommet marqué d'un triangle a la longueur $l = 10$.

La procédure de marquage pour un graphe de ce type consiste dans ce qui suit [12, 15] :

1. Associons à chaque sommet x_i un indice λ_i. Le sommet terminal x_0 a d'abord l'indice $\lambda_0 = 0$. Pour les autres sommets, posons en attendant $\lambda_i = \infty$ $(i \neq 0)$.

2. Cherchons un arc (x_i, x_j) tel que $\lambda_j - \lambda_i > l(x_i, x_j)$ et substituons à l'indice λ_j l'indice $\lambda'_j = \lambda_i + l(x_i, x_j) < \lambda_j$. Nous poursuivrons ce processus de substitution d'indices tant qu'il nous reste au moins un arc pour lequel λ_j peut être diminué.

Soulignons une propriété importante que possèdent les indices des sommets. Soit x_p un sommet arbitraire. Au cours du marquage décrit ci-dessus, l'indice λ_p décroît de façon monotone. Supposons que x_q soit le dernier sommet ayant contribué à la diminution de cet

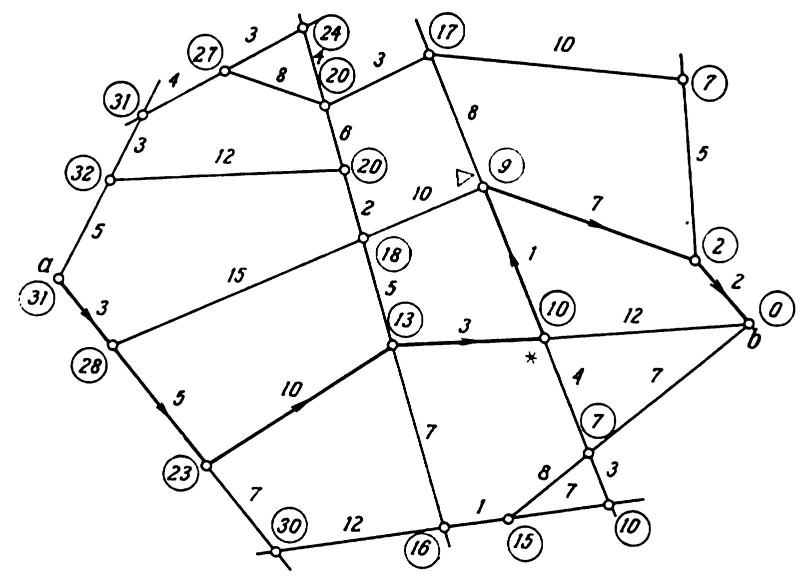


Fig. 2-6. Carte routière

indice. Alors $\lambda_p = \lambda_q + l(x_q, x_p)$. Donc, pour un sommet arbitraire x_p affecté d'indice λ_p, il y a un sommet x_q relié par une arête à x_p et tel que $\lambda_p - \lambda_q = l(x_q, x_p)$.

Cette propriété permet de formuler la règle suivante de recherche du plus court chemin.

Soit $x_n = a$ un sommet initial d'indice λ_n. On cherche un sommet x_{p_1} tel que $\lambda_n - \lambda_{p_1} = l(x_{p_1}, x_n)$. On cherche ensuite un sommet x_{p_2} tel que $\lambda_{p_1} - \lambda_{p_2} = l(x_{p_2}, x_{p_1})$, et ainsi de suite, jusqu'à gagner le sommet terminal $x_{p_{k+1}} = x_0 = b$. Le chemin $\mu_0 = (x_n, x_{p_1}, \ldots, x_{p_k}, x_0)$ de longueur λ_n est le plus court.

Pour le démontrer, considérons un chemin arbitraire de a à b: $\mu = (x_n, x_{k_1}, \ldots, x_{k_s}, x_0)$. Sa longueur est $l(\mu)$. Conformément à la règle de marquage, les inégalités suivantes seront vérifiées:

$$\left.\begin{array}{l} \lambda_n - \lambda_{k_1} \leqslant l(x_n, x_{k_1}); \\ \lambda_{k_1} - \lambda_{k_2} \leqslant l(x_{k_1}, x_{k_2}); \\ \ldots\ldots\ldots\ldots \\ \lambda_{k_s} - 0 \leqslant l(x_{k_s}, x_0). \end{array}\right\} \qquad (2\text{-}11)$$

Additionnant ces inégalités terme à terme, on trouve que pour tout chemin μ

$$\lambda_n - 0 \leqslant l(\mu). \qquad (2\text{-}12)$$

Parce qu'il vérifie la condition $\lambda_n = l(\mu_0)$, le chemin μ_0 est le plus court.

La méthode de recherche du plus court chemin est illustrée par la carte routière mise sous la forme d'un graphe à la figure 2-6. Les chiffres affectés aux arêtes indiquent le temps de parcours de chacun des chemins. Les indices des sommets fournissent le temps de parcours du sommet donné au sommet terminal.

d) Construction du graphe de longueur minimale

Nous passons maintenant à un problème de grande importance pratique, qui se présente sous la forme du problème du tracé des chemins. Il s'agit de relier plusieurs villes a, b, c, . . . entre elles par un réseau de routes. On connaît pour chaque couple de villes (x, y) le coût $l(x, y)$ de la construction de la route joignant ces villes. Le problème est de construire un réseau à minimum de frais.

Au lieu du réseau de routes, on peut considérer un réseau de lignes de transmission d'énergie électrique, un réseau de pipe-lines, etc. Si, dans le graphe représentant le réseau de routes, nous donnons à la quantité $l(x, y)$ l'appellation de longueur d'arête (x, y), nous aboutissons au problème de construction du graphe de longueur minimale. Aussi, au lieu du coût de construction, parlerons-nous par la suite de la longueur des arêtes du graphe.

S'il n'y a que trois sommets a, b, c, il suffit de construire entre eux une seule chaîne abc, acb ou bac ; si bc est l'arête la plus longue, on doit l'éliminer en construisant la chaîne bac.

Le graphe de longueur minimale est toujours un arbre, car, s'il contenait un cycle, il serait possible de supprimer une des arêtes de ce cycle sans que les sommets cessent d'être réunis. Par conséquent, si l'on veut réunir n sommets, on doit tracer $n - 1$ arêtes.

Montrons que le graphe de longueur minimale peut être construit à l'aide d'une règle suivante [14]. Relions tout d'abord l'un à l'autre les deux sommets les plus voisins: l'arête est u_1. A chaque pas suivant, ajoutons-y la plus courte des arêtes u_i qui, jointe aux arêtes déjà existantes, ne constitue avec celles-ci aucun cycle. S'il y a plusieurs arêtes de même longueur, choisissons-en une arbitrairement. A chaque arbre Q ainsi construit on donne le nom *d'arbre économique*. Sa longueur est égale à la somme des longueurs des arêtes isolées:

$$l(Q) = l(u_1) + \ldots + l(u_{n-1}). \qquad (2\text{-}13)$$

Montrons qu'aucun autre arbre reliant les mêmes sommets ne peut avoir la longueur inférieure à celle de l'arbre économique Q. Soient P l'arbre de longueur minimale reliant les sommets considérés, et Q un arbre économique quelconque. Supposons qu'on a numéroté les arêtes $u_1, u_2, \ldots, u_{n-1}$ en conservant la succession dans laquelle elles ont été tracées à la construction de Q, c.-à-d. qu'elles vérifient la condition $l(u_k) \leqslant l(u_{k+1})$. Si l'arbre P ne coïncide pas avec Q, cela veut dire que l'arbre Q possède une arête au moins n'appartenant pas à P. Soient $u_i = (a, b)$ une première arête de cette espèce et $L(a, b)$ la chaîne sur le graphe P, reliant les sommets a et b, comme par exemple sur la figure 2-7. Ajoutant l'arête u_i à P, on voit se former un cycle; or, puisque Q est démuni de cycles, ce cycle nouvellement formé doit contenir une arête au moins n'appartenant pas à Q. Supposons que ce soit par exemple u_i'. Supprimant cette arête, on obtient un arbre P' ayant le même nombre de sommets que P et dont la longueur est égale à

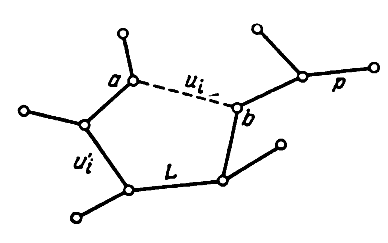


Fig. 2-7. Pour illustrer la construction de l'arbre de longueur minimale

$$l(P') = l(P) + l(u_i) - l(u_i'). \quad (2\text{-}14)$$

On sait que le graphe P a la longueur minimale. Donc

$$l(u_i) \geqslant l(u_i'). \quad (2\text{-}15)$$

Or, u_i était considérée comme la plus courte arête qui, jointe aux arêtes $u_1, u_2, \ldots, u_{i-1}$, ne donnait naissance à aucun cycle. Comme, en ajoutant u_i' à ces arêtes, on ne crée aucun cycle non plus, il vient

$$l(u_i) = l(u_i') \quad (2\text{-}16)$$

et P' a, de même que P, la longueur minimale. Or, P' a avec l'arbre économique Q une arête commune de plus que P. Reprenant cette opération plusieurs fois, on aboutit à un arbre de longueur minimale qui coïncidera avec Q. Par conséquent, Q est bien l'arbre de longueur minimale.

2-3. RÉSEAUX DE TRANSPORT

a) Notions principales

On entend par *réseau de transport* un graphe fini sans boucles caractérisé par:

1) un sommet et un seul x_0 tel que $\Gamma^{-1} x_0 = \varnothing$ (ce sommet s'appelle *entrée du réseau*);

2) un sommet et un seul z tel que $\Gamma z = \varnothing$ (il s'appelle *sortie du réseau*) ;

3) un nombre entier $c(u)$ rapporté à chaque arc u du graphe et appelé *capacité de l'arc u*.

A la notion de réseau de transport est étroitement liée celle de flot. Soit x un sommet quelconque. Désignons par U_x^- l'ensemble des arcs incidents à x vers l'intérieur, et par U_x^+ l'ensemble des arcs incidents à x vers l'extérieur. Le *flot dans le réseau* est la fonction $\varphi(u)$ vérifiant les conditions

$$0 \leqslant \varphi(u) \leqslant c(u), \quad u \in U; \tag{2-17}$$

$$\sum_{u \in U_x^-} \varphi(u) - \sum_{u \in U_x^+} \varphi(u) = 0;$$

$$x \neq x_0, \quad x \neq z. \tag{2-18}$$

La fonction $\varphi(u)$ peut être considérée comme la quantité de matière passant (à l'unité de temps) par l'arc $u = (x, y)$ de x vers y. Conformément à la condition (2-17), cette quantité ne peut être supérieure à la capacité $c(u)$ de l'arc. Suivant la condition (2-18), la quantité de matière arrivant à chaque sommet x autre que l'entrée x_0 et la sortie z doit être égale à la quantité de matière partant de ce sommet. Par conséquent, la matière ne peut s'accumuler nulle part dans le réseau, si ce n'est à l'entrée et à la sortie. Or, cela signifie que le flot sortant du sommet d'entrée x_0 est exactement égal au flot entrant dans le sommet de sortie z:

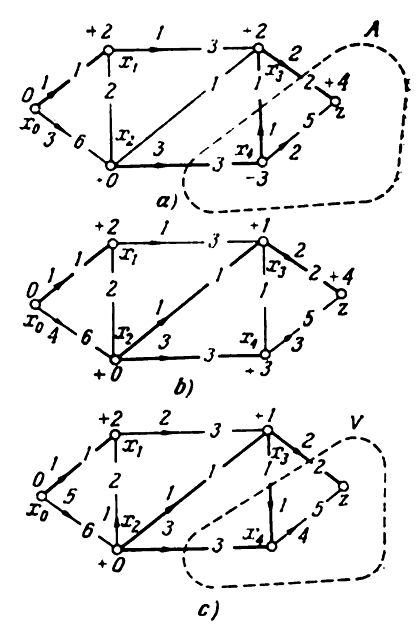


Fig. 2-8. Répartition des flots dans le réseau de transport

$$\sum_{u \in U_{x_0}^+} \varphi(u) = \sum_{u \in U_z^-} \varphi(u) = \varphi_z. \tag{2-19}$$

La quantité $\varphi(z)$ porte le nom de *valeur du flot* dans le réseau de transport.

Un exemple du réseau de transport est donné sur la figure 2-8. Les chiffres placées dans les coupures des arcs indiquent la capacité de l'arc. Les flèches indiquent la direction des flots, et les chiffres affectant les flèches, la valeur du flot. C'est par analyse d'un réseau de transport qu'on trouve la solution de nombreux problèmes de

planification des livraisons, de distribution des marchandises entre les consommateurs, etc.

L'étude de la répartition du flot dans le réseau de transport se trouve grandement facilitée après l'introduction de la notion de coupe du réseau. Soit $A \subset X$ un ensemble vérifiant les conditions

$$x_0 \notin A, \quad z \in A. \tag{2-20}$$

Désignons par U_A^- et U_A^+ les ensembles des arcs incidents à A respectivement vers l'intérieur et vers l'extérieur. C'est à la totalité des arcs $U_A = U_A^- \cup U_A^+$ qu'on donne le nom de *coupe A* du réseau de transport. A la figure 2-8 on voit un exemple de coupe.

Du moment que toute particule de matière allant de x_0 vers z passera sûrement par l'un quelconque des arcs de la coupe, le flot total dans la coupe sera égal à la valeur du flot dans le réseau, c.-à-d. que pour toute coupe A on a

$$\varphi_z = \sum_{u \in U_A^-} \varphi(u) - \sum_{u \in U_A^+} \varphi(u). \tag{2-21}$$

La *capacité de la coupe A* est la somme des capacités des arcs faisant partie de la coupe:

$$c(A) = \sum_{u \in U_A^-} c(u). \tag{2-22}$$

Puisque pour tout arc $\varphi(u) \leqslant c(u)$, il résulte de (2-21) et de (2-22) que

$$\varphi_z \leqslant c(A). \tag{2-23}$$

b) Problème du flot maximal

Le problème du flot maximal dans un réseau de transport consiste dans la recherche de la valeur du plus grand flot dont est capable le réseau et de la répartition de ce flot suivant les arcs du réseau, la configuration du réseau et les capacités de ses arcs étant connues.

Lemme. *Si pour une valeur quelconque du flot dans un réseau de transport φ_z et une coupe donnée V la condition $\varphi_z = c(V)$ est remplie, le flot φ_z est maximal et la coupe V a la capacité minimale.*

D é m o n s t r a t i o n. On a vu que la valeur du flot φ_z pour toute coupe A doit vérifier la relation (2-23). Désignons par V la coupe de capacité minimale

$$c(V) = \min_A c(A). \tag{2-24}$$

La valeur du flot φ_z étant la même pour toute coupe du réseau de transport, l'accroissement de la valeur du flot φ_z n'est possible

que jusqu'au moment où il atteint la valeur $c(V)$. Donc, la valeur du flot

$$\varphi_z = c(V) \tag{2-25}$$

définit justement le flot maximal dans le réseau de transport.

Le lemme que l'on vient de considérer ne fournit cependant pas de méthode pratiquement acceptable pour la recherche du flot maximal. Pour formuler cette méthode, faisons intervenir encore quelques définitions auxiliaires.

Nous dirons d'un arc u qu'il est *saturé* quand $\varphi(u) = c(u)$, et d'un flot φ_z qu'il est *complet* lorsque tout chemin de x_0 à z passe au moins par un arc saturé.

Le flot complet dans le réseau de transport donné n'est pas une quantité bien déterminée: il dépend de la direction des flots dans les différents arcs. Par exemple, on voit sur la figure 2-8,*a* et *c* deux répartitions différentes du flot dans le même réseau de transport. Les arcs saturés sont représentés par des lignes grasses. Dans les deux cas les flots sont complets bien que leurs valeurs soient différentes.

L'algorithme pour la recherche du flot maximal, proposé par Ford et Fulkerson [15], consiste à faire accroître progressivement le flot φ_z jusqu'à ce qu'il devienne maximal. Ce faisant, on admet que les capacités des arcs $c(u)$ s'expriment par des nombres entiers, de sorte que les flots dans les arcs seront aussi des nombres entiers. On cherche le flot maximal en deux étapes.

1. *Détermination du flot complet.* Supposons que $\varphi(u)$ soit une certaine répartition du flot dans les arcs du réseau de transport. Cherchant un chemin μ de x_0 à z tel que tous ses arcs soient insaturés, on admet que

$$\varphi'(u) = \begin{cases} \varphi(u)+1 & \text{quand } u \in \mu; \\ \varphi(u) & \text{quand } u \notin \mu. \end{cases} \tag{2-26}$$

Le flot φ_z varie alors jusqu'à atteindre la valeur $\varphi'_z = \varphi_z + 1 > \varphi_z$. On augmente de cette manière progressive le flot φ_z jusqu'à ce qu'il devienne complet.

Exemple 2-3. Cherchons le flot complet dans le réseau de transport de la figure 2-8,*a*. On considère successivement quelques chemins en marquant les arcs saturés en trait gras:

$\mu_1 = (x_0, x_1, x_3, z)$, $\varphi(\mu_1) = 1$; arc saturé (x_0, x_1);

$\mu_2 = (x_0, x_2, x_4, x_3, z)$, $\varphi(\mu_2) = 1$; arcs saturés (x_4, x_3) et (x_3, z);

$\mu_3 = (x_0, x_2, x_4, z)$; pour que ce chemin soit saturé, on peut prendre $\varphi(\mu_3) = 2$; arc saturé (x_2, x_4).

Il est facile de voir qu'il n'y a plus aucun chemin de x_0 à z qui contienne des arcs insaturés. Donc, le flot complet

$$\varphi_z = \varphi(\mu_1) + \varphi(\mu_2) + \varphi(\mu_3) = 4.$$

2. *Détermination du flot maximal.* Soient φ_z le flot complet et $\varphi(x, y)$ le flot dans l'arc $u = (x, y)$ dirigé du sommet x vers le sommet y. Le processus d'accroissement de φ_z se ramène à attribuer aux sommets du graphe des indices indiquant le chemin qui rend possible l'accroissement du flot. On doit préalablement numéroter tous les sommets du graphe.

Donnons à x_0 l'indice 0. Si x_i est un sommet déjà marqué, on attribue l'indice $+i$ à tous les sommets non marqués auxquels aboutissent des arcs insaturés incidents à x_i vers l'extérieur, c.-à-d. aux sommets y pour lesquels

$$(x_i, y) \in U \quad \text{et} \quad \varphi(x_i, y) < c(x_i, y) \tag{2-27}$$

et l'indice $-i$, à tous les sommets non marqués d'où partent des arcs incidents à x_i vers l'intérieur, c.-à-d. aux sommets y pour lesquels

$$(y, x_i) \in U \quad \text{et} \quad \varphi(y, x_i) > 0. \tag{2-28}$$

Si, à la suite de ce processus, on constate que le sommet z est marqué, on peut être sûr de trouver entre x_0 et z une chaîne dont tous les sommets sont distincts et (au signe près) sont porteurs des numéros des sommets précédents. Faisons accroître d'une unité le flot dans toutes les arêtes de cette chaîne dans la direction de x_0 vers z en admettant que

$$\varphi'(u) = \varphi(u) \text{ si } u \notin \mu;$$
$$\varphi'(u) = \varphi(u) + 1 \text{ si } u \in \mu$$

et qu'en allant de x_0 vers z on parcourt l'arc u dans le sens direct de son orientation;

$$\varphi'(u) = \varphi(u) - 1 \text{ si } u \in \mu$$

et qu'en allant de x_0 à z on parcourt l'arc u dans le sens inverse de son orientation.

A la suite de ce processus on obtient un nouveau flot dans le réseau $\varphi'_z = \varphi_z + 1$, ce qui signifie que la valeur du flot croît. Ensuite, on reprend ce processus.

S'il s'avère impossible d'accroître un flot φ_z^0 par la méthode proposée, c.-à-d. d'attribuer un indice au sommet z, alors φ_z^0 est le flot maximal dans le réseau. En effet, soit V l'ensemble des sommets non marqués, dont fait partie bien sûr le sommet z. Par conséquent, V est une coupe telle qu'aucun arc n'y est incident vers l'extérieur (dans le cas contraire certains sommets de cette coupe seraient marqués avec les indices négatifs) et que tous les arcs incidents à cette coupe vers l'intérieur sont saturés:

$$\left.\begin{aligned} &U_V^- = U_V; \quad U_V^+ = \varnothing; \\ &\varphi(u) = c(u) \quad \text{pour} \quad u \in U_V^-. \end{aligned}\right\} \tag{2-29}$$

On a en outre

$$\varphi_z^0 = \sum_{u \in U_V^-} \varphi(u) - \sum_{u \in U_V^+} \varphi(u) = \sum_{u \in U_V^-} c(u) - 0 = c(V). \quad (2\text{-}30)$$

En vertu du lemme démontré ci-dessus, φ_z^0 est le flot maximal et V la coupe de capacité minimale.

Exemple 2-4. On a marqué les sommets d'un réseau de transport comme il est montré sur la figure 2-8,*a*. Le sommet z est affecté de l'indice $+4$. La suite décroissante d'indices $+4$, -3, $+2$, $+0$ définit la chaîne $\mu = (x_0, x_2, x_3, x_4, z)$ dans laquelle il convient d'augmenter d'une unité le flot de x_0 jusqu'à z. Ce faisant, on obtient la répartition du flot montrée sur la figure 2-8,*b*. Reprenant la procédure de marquage sur cette figure, on trouve la chaîne $\mu = (x_0, x_2, x_1, x_3, x_4, z)$ dont le flot doit lui aussi être augmenté d'une unité. La répartition du flot qui en résulte est montrée sur la figure 2-8,*c*. Deux sommets x_4 et z ne sont pas marqués sur cette figure.

Ainsi donc, la répartition du flot obtenue assure le flot maximal φ_z^0 dans le réseau de transport considéré et l'ensemble $V = \{x_4, z\}$ définit la coupe de capacité minimale. On trouve la valeur du flot φ_z^0 après avoir déterminé la capacité de la coupe V:

$$\varphi_z^0 = c(x_2, x_4) + c(x_3, x_4) + c(x_3, z) = 3 + 1 + 2 = 6.$$

L'augmentation du flot maximal dans le réseau de transport peut être réalisée en augmentant la capacité de l'un quelconque des arcs incidents à la coupe V vers l'intérieur.

c) Problème de transport

A côté du problème de recherche du flot maximal, une grande importance pratique est attachée au problème de la répartition la plus économique d'un flot suivant les arcs du réseau de transport, désigné sous le terme de *problème de transport*. Le réseau de transport représentant dans bien des cas le schéma d'organisation des transports d'une marchandise quelconque, la solution du problème de transport permet de déterminer le plan le plus rationnel de transport, c.-à-d. la répartition des itinéraires qui assure, par exemple, le coût minimal du transport, ou bien la livraison des marchandises au client en un temps aussi court que possible. Le problème du premier genre est dit problème de transport à minimum de frais, et celui du deuxième genre, problème de transport à minimum de temps.

Pour faciliter l'exposé, adoptons les désignations $c_{ij} = c(x_i, x_j)$ pour la capacité de l'arc (x_i, x_j), et $d_{ij} = d(x_i, x_j)$ pour le coût de passage de l'unité de flot suivant l'arc (x_i, x_j).

Le *problème de transport à minimum de frais* se formule, en termes de la théorie des graphes, de la manière suivante.

Soient donnés un réseau de transport caractérisé par le flot maximal φ_z^0 et un flot $\varphi_z \leqslant \varphi_z^0$ qu'il s'agit de faire passer par ce réseau. On demande de savoir une répartition du flot φ_z suivant les arcs du réseau telle qu'elle assure le coût minimal de passage du flot. On demande en outre que soit vérifiée, pour chaque arc, la relation

$\varphi(x_i, x_j) \leqslant c_{ij}$ et que le coût de passage du flot $\varphi(x_i, x_j)$ suivant l'arc (x_i, x_j) soit égal à $d_{ij}\,\varphi(x_i, x_j)$.

Pour résoudre ce problème, considérons les quantités d_{ij} comme longueurs des arcs correspondants. Dans ce cas, le coût de passage du flot φ par un chemin quelconque μ entre x_0 et z sera égal au produit de la longueur du chemin par la valeur du flot φ et le problème de minimisation du coût de passage du flot sera ramené à la recherche du chemin de longueur minimale dans le graphe de x_0 à z, problème auquel on a eu affaire un peu plus haut. Dans le cas où aucune restriction n'est imposée à la capacité des arcs, le plus court chemin est justement celui qui assure le coût minimal de passage du flot.

S'il y a des restrictions imposées à la capacité des arcs, le problème est résolu en quelques étapes, par la recherche des flots partiels à chaque étape. Dans sa forme générale, le mode de résolution du problème consiste dans ce qui suit.

Dans le graphe $G_1 = (X, \Gamma)$ représentant le réseau de transport aux arcs de longueur $d_{ij} = l(x_i, x_j)$, proposons-nous de chercher le chemin le plus court μ_1 de x_0 à z. Soit c_1 la capacité du chemin μ_1. On fait passer par ce chemin un flot

$$\varphi_1 = \begin{cases} \varphi_z & \text{si } \varphi_z \leqslant c_1; \\ c_1 & \text{si } \varphi_z > c_1. \end{cases} \tag{2-31}$$

Si $\varphi_z \leqslant c_1$, le problème est résolu et le chemin μ_1 est le plus économique pour le flot φ_z.

Si $\varphi_z > c_1$, le flot φ_1 est considéré comme un flot partiel et on passe au graphe G_2 obtenu à partir de G_1 en remplaçant les capacités des arcs c_{ij} par c'_{ij} tirées de la relation

$$c'_{ij} = \begin{cases} c_{ij} - c_1 & \text{pour } u \in \mu_1; \\ c_{ij} & \text{pour } u \notin \mu_1. \end{cases} \tag{2-32}$$

Les arcs pour lesquels $c'_{ij} = 0$ ne sont pas pris en considération. On admet que le flot dont on cherche la répartition dans le graphe G_2 est égal à

$$\varphi'_z = \varphi_z - \varphi_1. \tag{2-33}$$

Maintenant, on aborde de nouveau le problème initial de recherche de la répartition la plus économique du flot φ'_z mais déjà dans le graphe G_2. La solution de ce problème fournit le chemin μ_2 de capacité c_2 suivant lequel on fait passer le flot partiel

$$\varphi_2 = \begin{cases} \varphi'_z & \text{si } \varphi'_z \leqslant c_2; \\ c_2 & \text{si } \varphi'_z > c_2. \end{cases} \tag{2-34}$$

Si $\varphi'_z \leqslant c_2$, le problème est résolu, et la répartition la plus économique des flots dans le graphe G_1 consiste à canaliser le flot φ_1 suivant l'itinéraire μ_1 et le flot φ_2 suivant l'itinéraire μ_2.

Si $\varphi_z' > c_2$, il convient de passer à un nouveau graphe G_3 et de chercher un nouveau flot partiel φ_3. Ce processus est renouvelé jusqu'au moment où la somme des flots partiels devienne égale à φ_z. Ces flots partiels passés à travers le graphe G_1 représentent justement la répartition la plus économique du flot φ_z.

Pour illustrer la méthode décrite, considérons une variante largement répandue du problème de transport à minimum de frais.

Une marchandise homogène est stockée aux stations $x_1, \ldots, x_m$ en quantités $a_1, \ldots, a_m$. On demande de transporter cette marchandise aux stations $y_1, \ldots, y_r$ en quantités $b_1, \ldots, b_r$. On suppose que la quantité totale de marchandise à expédier soit égale aux réserves disponibles:

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{r} b_j. \qquad (2\text{-}35)$$

Le coût de transport de la marchandise de la station x_i à la station y_j est égal à d_{ij}. On demande de fixer les itinéraires de transport les plus économiques. Il est commode d'écrire les données initiales sous la forme d'une table (voir tableau 2-1).

Tableau 2-1

Problème de transport

a_i	b_j		
	b_1	...	b_r
a_1	d_{11}	...	d_{1r}
...	...	...	...
a_m	d_{m1}	...	d_{mr}

Tableau 2-2

Données initiales au problème de transport

a_i	b_j			
	5	10	20	15
10	8	3	5	2
15	4	1	6	7
25	1	9	4	3

Le réseau de transport correspondant à ce problème est construit de la façon suivante. On relie l'entrée x_0 à chacun des sommets x_i par un arc de capacité $c(x_0, x_i) = a_i$. On relie ensuite chacun des sommets y_j à la sortie z par un arc de capacité $c(y_j, z) = b_j$. Le coût de passage du flot suivant les arcs (x_0, x_i) et (y_j, z) est considéré comme nul. Enfin, chaque sommet x_i est relié à chaque sommet y_j par un arc de capacité infinie; le coût de passage de l'unité de flot par cet arc est égal à d_{ij}. Ceci fait, on applique au réseau de transport ainsi constitué la méthode décrite ci-dessus.

Exemple 2-5. On demande de trouver les itinéraires les plus économiques pour le problème de transport résumé au tableau 2-2. Ce tableau correspond au

Tableau 2-3

Répartition des flots partiels dans le problème de transport à minimum de frais

k	Itinéraire (x_i, y_j)	Flot partiel φ_k	d_{ij}	Coût de transport $d_{ij}\varphi_k$
1	(x_3, y_1)	5	1	5
2	(x_2, y_2)	10	1	10
3	(x_1, y_4)	10	2	20
4	(x_3, y_4)	5	3	15
5	(x_3, y_3)	15	4	60
6	(x_2, y_3)	5	6	30
—	—	50	—	140

schéma ferroviaire (fig. 2-9) reliant les usines d'éléments de construction aux consommateurs de ces éléments (chantiers). Le réseau de transport correspondant aux données du tableau 2-2 est représenté sur la figure 2-10.

Appliquant la méthode décrite ci-dessus, on trouve les flots partiels et les itinéraires, énumérés dans le tableau 2-3 dans l'ordre de leur obtention. Le plan des transports correspondant à ce tableau est montré aussi sur le schéma ferroviaire.

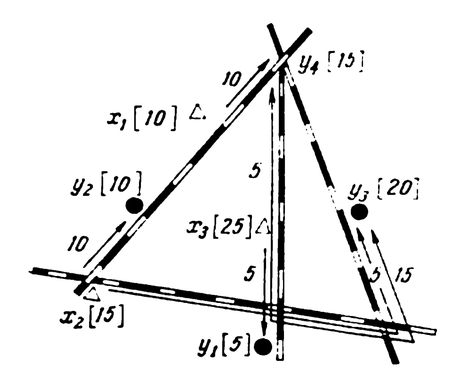


Fig. 2-9. Schéma ferroviaire

Il est facile de voir que dans le cas général le coût de transport des marchandises suivant le réseau du type considéré se définit par l'expression

$$\sum_{i=1}^{m} \sum_{j=1}^{r} d_{ij}\varphi(x_i, y_j). \quad (2\text{-}36)$$

On en conclut que la méthode proposée de résolution du problème de transport fournit en principe le mode d'obtention des valeurs des flots partiels $\varphi(x_i, y_j)$ minimisant la somme mentionnée. Ce mode n'est pas unique: dans le chapitre consacré à la programmation linéaire nous aurons l'occasion d'étudier quelques autres méthodes de résolution de problèmes analogues.

Le *problème de transport à minimum de temps* sera examiné dans le réseau de transport résumé au tableau 2-2; les quantités d_{ij} seront considérées désormais comme le temps nécessaire au transport de la marchandise du point x_i au point y_j et notées par t_{ij} dans le texte qui suit. De pareils problèmes se présentent lorsqu'il s'agit de transporter des produits à durée de conservation limitée, d'expédier des moyens de secours dans les zones sinistrées, d'acheminer les céréales

nouvellement récoltées aux moulins, etc. Dans tous ces cas on exige de faire parvenir toutes les marchandises à destination en un temps aussi court que possible.

Considérons le mode de résolution général de ce problème. Supposons que l'on ait trouvé, par une méthode quelconque, une répartition du flot φ_z dans le graphe G représentant le réseau de transport

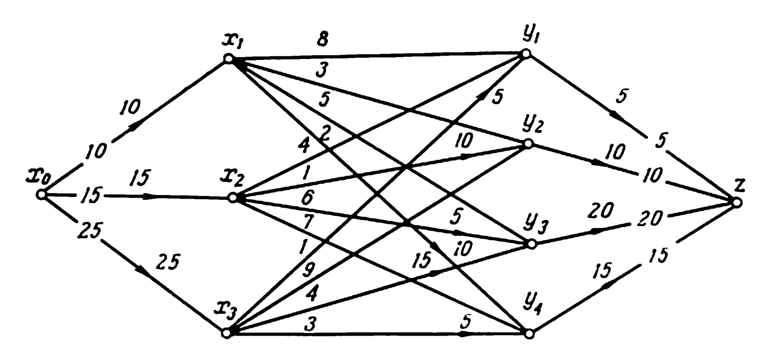


Fig. 2-10. Solution du problème de transport à minimum de frais

considéré. Délimitons sur G un graphe partiel G' ne comprenant que les arcs par lesquels est transmis le flot φ_z. Soient μ un chemin allant de x_0 à z, et t_μ, le temps de parcours du flot suivant ce chemin. Il est évident que le temps nécessaire au transport de toutes les marchandises de x_0 à z est défini par le chemin parcouru par le flot à maximum de temps, car le transport des marchandises suivant les autres chemins s'effectue plus vite. Donc, le temps T nécessaire pour transporter toutes les marchandises est égal à

$$T = \max_{\mu \in G'} t_\mu. \tag{2-37}$$

De cette façon, la solution du problème de transport à minimum de temps s'obtient en dégageant du graphe G un graphe partiel G' tel qu'il puisse être parcouru par le flot φ_z tout entier et dans lequel la durée du chemin le plus prolongé soit minimale par rapport à tous les autres graphes de cette nature. Il peut se trouver d'ailleurs que la solution trouvée ci-dessus en recherchant le minimum de frais et qui minimise la quantité définie par l'expression (2-36) ne sera point optimale au point de vue de l'impératif du temps minimal.

La résolution du problème posé se ramène à perfectionner pas à pas le graphe G' en supprimant dans celui-ci les chemins les plus prolongés et en y ajoutant des chemins plus courts mais non utilisés auparavant, et à répartir convenablement le flot φ_z.

Revenons au même exemple. En qualité de première approximation de la solution optimale, adoptons la solution trouvée en recherchant le minimum de frais. On voit à la figure 2-10 la répartition des flots pour ce cas. On remarque que le temps de parcours du flot suivant l'itinéraire le plus prolongé (x_2, y_3)

est égal à 6. Or, il y a des itinéraires moins prolongés (x_1, y_2), (x_2, y_1) (x_1, y_3) qui n'ont pas été utilisés. Il est donc possible que l'un quelconque de ces itinéraires puisse être substitué à (x_2, y_3).

Construisons un graphe partiel G' ne comprenant que les arcs à $t_{ij} < 6$ du graphe G; la répartition du flot restera la même que dans G. Le graphe G'

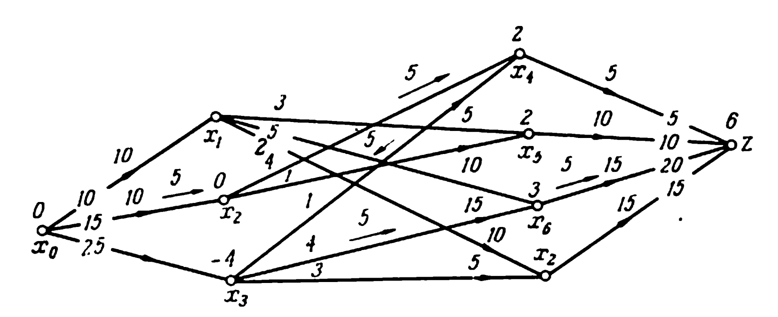


Fig. 2-11. Solution du problème de transport à minimum de temps

est représenté sur la figure 2-11 où les sommets y_j sont désignés, pour plus de commodité, par $x_{m+j} = x_{3+j}$. Le flot φ'_z dans ce graphe est égal à 45 unités, donc inférieur au flot initial $\varphi_z = 50$ unités. Ce flot est complet, car tous les chemins dans G' allant de x_0 à z contiennent des arcs saturés. Il n'en est pas moins possible qu'il ne soit pas maximal.

Si le flot maximal dans le graphe G' est égal à φ_z, on peut être sûr de trouver une répartition de ce flot telle que le chemin le plus prolongé prendra $t_\mu < 6$ unités. Il reste maintenant de trouver le flot maximal dans G'.

On voit sur la figure 2-11 les sommets du graphe G' affectés d'indices d'après la règle énoncée au paragraphe 2-3. Les indices témoignent de l'existence d'un chemin $(x_0, x_2, x_4, x_3\ x_6, z)$ qui permet d'accroître le flot de x_0 jusqu'à z de 5 unités. Ce flot supplémentaire est montré par des flèches. On voit que le flot maximal dans G' est égal à $\varphi_z=50$ unités et l'itinéraire le plus prolongé dure 4 unités de temps. Il n'est pas possible de diminuer davantage la durée de passage du flot, car, parmi les itinéraires aboutissant à y_3 (x_6), aucun n'a le temps de parcours inférieur à 4 unités.

La répartition définitive des flots partiels suivant les itinéraires, fournissant la solution du problème de transport à minimum de temps, est résumée au tableau 2-4.

Tableau 2-4

Répartition des flots partiels dans le problème de transport à minimum de temps

Itinéraire	Flot partiel	Temps de parcours du flot
(x_2, y_2)	10	1
(x_1, y_4)	10	2
(x_3, y_4)	5	3
(x_2, y_1)	5	4
(x_3, y_3)	20	4
—	$\varphi_z=50$	$t_{max}=4$

CHAPITRE 3

ESPACES MULTIDIMENSIONNELS

3-1. ESPACES MÉTRIQUES ET DISTANCES

a) Notion de distance

Les ensembles ne peuvent être considérés comme collections de certains objets, ainsi qu'il a été fait au chapitre 1, que d'une façon assez limitée, car tous les objets matériels de la nature se trouvent en liaison réciproque et en interaction. Aussi est-il nécessaire de définir l'ensemble au point de vue de l'établissement de telles ou telles relations entre ses éléments.

On dit que l'ensemble est muni d'une *structure* si certaines relations sont établies entre ses éléments et que certaines opérations soient définies sur ces derniers. Un ensemble muni d'une structure porte le nom d'*espace*.

Fig. 3-1. Illustration de l'inégalité triangulaire

Nous allons commencer l'étude des espaces par l'espace le plus élémentaire, dit *espace métrique*. Pour définir celui-ci, on doit introduire la notion de *distance* des éléments d'un ensemble.

La notion de distance est une notion bien familière à l'homme, qui associe cette notion à la manière de situer les objets dans l'espace et qui entend par distance la mesure d'éloignement des objets entre eux. Généralement, la distance $d(M, N)$ entre deux points M et N a pour mesure la longueur du segment reliant ces points (fig. 3-1). Or, bien souvent, une telle définition de la distance s'avère insuffisante. Même dans la vie quotidienne, par exemple, la distance entre deux villes ne peut être définie de façon univoque (distance par chemin de fer, par voie d'eau, etc.). Dans une ville partagée en pâtés de maisons, comme il est montré sur la figure 3-2, il n'y a aucun sens à relier les points M et N par un segment de droite pour en mesurer la distance, du moment qu'on ne marche que dans les rues.

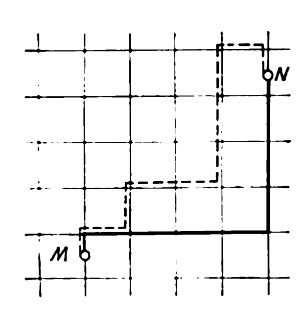


Fig. 3-2. Détermination de la distance dans une ville divisée en pâtés de maisons

D'autre part, il arrive souvent qu'en disant « loin », « éloigné », on ne pense pas à l'espace dans le sens habituel du mot mais on parle, par exemple, du temps (« un passé lointain »...). Si un certain système

est susceptible de prendre successivement des états A_1, A_2, . . ., A_n, l'éloignement de l'état A_k relativement à l'état A_j peut avoir pour mesure le nombre d'états que le système doit prendre pour passer de l'état A_j à l'état A_k. La distance est ici la mesure d'éloignement des états du système entre eux. Or, si l'éloignement est considéré comme une propriété de l'espace, on est amené à considérer non plus l'espace ordinaire à trois dimensions mais un espace de nature différente; on peut l'appeler par exemple espace des états.

Les exemples ci-dessus montrent qu'il doit exister une définition générale de la distance comme mesure d'éloignement d'objets et donc une définition générale de l'espace dans lequel ces objets existent, ces notions pouvant avoir une signification différente dans les différentes situations concrètes. Puisque les collections d'objets divers représentent des ensembles, les notions d'espace et de distance doivent être liées à celle d'ensemble.

b) Définition de l'espace métrique

Soit X un ensemble quelconque. La notion de distance entre les éléments de X résulte de la généralisation des propriétés fondamentales considérées intuitivement comme inhérentes à la distance et qu'on illustre clairement sur la figure 3-1.

Faisons correspondre à tout couple d'éléments de X un nombre réel non négatif $d \geqslant 0$. Ce nombre est appelé *distance* ou *métrique* de X s'il satisfait, pour tous les x, y, $z \in X$, aux trois conditions (axiomes) suivantes:

1) $d(x, y) = 0$ si et seulement si $x = y$ (axiome d'identité);
2) $d(x, y) = d(y, x)$ (axiome de symétrie);
3) $d(x, y) \leqslant d(x, z) + d(z, y)$ quels que soient x, y, $z \in X$ (inégalité triangulaire).

L'*espace métrique* est le couple (X, d), c.-à-d. l'ensemble X à métrique d définie sur cet ensemble. Les éléments de l'ensemble X sont dits les *points* de l'espace métrique (X, d).

Il découle de cette définition qu'un ensemble X ne devient un espace métrique que par l'introduction d'une métrique correspondante $d(x, y)$. Introduisant dans un seul et même ensemble X diverses métriques, on obtient plusieurs espaces différents. C'est ainsi que les espaces représentés sur les figures 3-1 et 3-2 ont pour éléments les ensembles des points du plan mais possèdent des métriques différentes.

c) Exemples d'espaces métriques

1. Soient x, y des éléments arbitraires de l'ensemble R des nombres réels. On peut faire de l'ensemble R un espace métrique en définissant la distance entre x et y par la formule

$$d(x, y) = |x - y|. \qquad (3\text{-}1)$$

C'est avec cette formule qu'on trouve la distance séparant des points de l'axe réel, qui est l'exemple le plus élémentaire d'espace métrique.

2. Considérons l'ensemble R^n qui a pour éléments des n-uplets ordonnés des nombres réels de la forme $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$. Il existe plusieurs procédés pour faire de R^n un espace métrique. La distance entre les points x et y est définie le plus souvent par la formule

$$d_2(x, y) = \{\sum_{i=1}^{n} |x_i - y_i|^2\}^{1/2}. \tag{3-2}$$

Pour $n = 2, 3$ cette définition se confond avec la notion géométrique habituelle de la distance. Les propriétés 1, 2, 3 pour cette distance sont évidentes (fig. 3-1).

On appelle la métrique $d_2(x, y)$ *métrique euclidienne*, et l'espace R^n muni d'une telle métrique et noté E_n, *espace euclidien.*

3. Pour l'ensemble R^n la distance peut aussi être définie autrement, par exemple:

$$d_1(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{3-3}$$

ou

$$d_\infty(x, y) = \max(|x_1 - y_1|, \ldots, |x_n - y_n|). \tag{3-4}$$

Il est facile de voir que les métriques d_2, d_1, d_∞ sont des cas particuliers de la métrique

$$d_p(x, y) = \{\sum_{i=1}^{n} |x_i - y_i|^p\}^{1/p}$$

et s'obtiennent respectivement pour $p = 2$, $p = 1$ et $p = \infty$.

Les propriétés 1 et 2 pour les métriques (3-3) et (3-4) sont immédiates. Pour démontrer la propriété 3, prenons en outre un point $z = (z_1, \ldots, z_n) \in R^n$. La distance $d_1(x, y)$ s'écrit:

$$d_1(x, y) = \sum_{i=1}^{n} |x_i - z_i + z_i - y_i| \leqslant$$
$$\leqslant \sum_{i=1}^{n} (|x_i - z_i| + |z_i - y_i|) = d(x, z) + d(z, y).$$

Pour la distance $d_\infty(x, y)$ la propriété 3 se vérifie comme suit. Supposons que $|x_k - y_k|$ soit la plus grande des différences correspondantes des points x et y. Alors

$$d_\infty(x, y) = |x_k - y_k| = |x_k - z_k + z_k - y_k| \leqslant$$
$$\leqslant |x_k - z_k| + |z_k - y_k|.$$

Il est évident que

$$|x_k - z_k| \leqslant \max(|x_1 - z_1|, \ldots, |x_n - z_n|) = d_\infty(x, z);$$
$$|z_k - y_k| \leqslant \max(|z_1 - y_1|, \ldots, |z_n - y_n|) = d_\infty(z, y).$$

Donc,

$$d_\infty(x, y) \leqslant d_\infty(x, z) + d_\infty(z, y).$$

4. Considérons l'ensemble de toutes les fonctions du temps possibles qui sont continues sur l'intervalle $a \leqslant t \leqslant b$. Soient $x(t)$ et $y(t)$ deux fonctions de cette espèce. Leur distance s'exprime par la relation

$$d(x, y) = \max_{a \leqslant t \leqslant b} |x(t) - y(t)|, \tag{3-5}$$

qui satisfait à toutes les propriétés d'une métrique (cela est facile à vérifier). L'espace doté d'une telle métrique se note $C_{[a, b]}$.

Pour illustrer les larges possibilités d'emploi des notions nouvellement introduites, il y a lieu de considérer un espace dont l'importance en cybernétique est primordiale, à savoir l'espace des messages.

3-2. INTERPRÉTATION GÉOMÉTRIQUE DES SIGNAUX ET DES MESSAGES

a) Espace des messages

Nous avons vu en Introduction que les messages sont transmis par les canaux de communication au moyen d'un alphabet $\mathfrak{A}$ constitué par des symboles en nombre fini. Les messages représentent diverses suites de symboles alphabétiques. Le nombre de symboles dans le message est la longueur du message.

Considérons un cas où l'on se sert d'un alphabet $\mathfrak{A}_m$ comprenant m symboles pour transmettre des messages de longueur n. On range parmi ces derniers aussi bien des messages plus brefs, à condition de les compléter jusqu'à n par addition d'un symbole déterminé; en notation binaire, ce sera le signe zéro. L'ensemble réunissant tous ces messages peut être assimilé à un espace métrique par l'introduction de la notion de distance des messages.

Convenons d'appeler distance $d(x, y)$ de deux messages x et y le nombre de positions dans lesquelles les messages x et y ont des symboles différents. L'espace métrique ainsi obtenu sera noté E $(n, \mathfrak{A}_m)$ et sera appelé *espace des messages.*

Exemple 3-1. $\mathfrak{A}_m$ est l'alphabet français, $n = 7$, $x =$ (chapeau), $y =$ (château). La quatrième lettre ne coïncide pas. Donc, $d(x, y) = 1$.

Exemple 3-2. Soit $\mathfrak{A} = \mathfrak{A}_2 = \{0, 1\}$ l'alphabet binaire, $n = 10$, $x = 0100111010$, $y = 0010110010$. Le deuxième, le troisième et le septième signe ne coïncident pas. Donc, $d(x, y) = 3$.

La définition de la distance $d(x, y)$ dans l'espace des messages, telle qu'on vient de la donner, satisfait à toutes les propriétés de la distance. L'axiome d'identité et l'axiome de symétrie sont immédiats. Montrons qu'il en est de même de l'axiome d'inégalité triangulaire.

Soient trois messages x, y et z de longueur n. Soient x_k, y_k et z_k les symboles de ces messages en k-ième position. Il est naturel que si $x_k = y_k$ et si $y_k = z_k$, on ait $x_k = z_k$, c.-à-d. que si en position quelconque les symboles des messages x et y et des messages y et z coïncident, ils coïncident aussi dans cette position pour les messages x et z. Donc, pour les messages x et z les symboles non coïncidants ne peuvent se trouver que là où les symboles ne coïncident pas soit pour les messages x et y, soit pour les messages y et z. Or, cela signifie que le nombre total de symboles qui ne coïncident pas dans les messages x et z ne peut être supérieur à la somme des nombres de symboles qui ne coïncident pas dans x et y et dans y et z.

Par la suite, dans le souci de rendre notre exposé plus simple et clair, nous nous bornerons au seul alphabet binaire $\mathfrak{A} = \mathfrak{A}_2 = \{0, 1\}$. On peut le faire sans restreindre la généralité des raisonnements, puisque les messages écrits par des symboles d'un alphabet quelconque peuvent toujours être transcrits au moyen de symboles d'un autre alphabet. Soit, par exemple, un alphabet $\mathfrak{A}_m$ à m symboles. Affectons chaque symbole d'un numéro d'ordre de 0 à $m - 1$. Substituant maintenant dans notre message aux symboles de l'alphabet $\mathfrak{A}_m$ leurs numéros d'ordre en numération binaire, nous obtiendrons le même message écrit à l'aide de l'alphabet binaire.

b) Codes à détection et à correction d'erreurs

Au cours de la transmission par le canal de communication, le message peut être déformé. Dans le cas de l'alphabet binaire, la déformation se manifeste par l'apparition de quelques zéros à la place des unités et de quelques unités à la place des zéros. La question se pose de savoir s'il est possible d'inventer des codes tels qu'ils puissent signaler les déformations subies par le message au cours de la transmission ou même reconstituer les valeurs des bits mutilés?

Considérons le cas où, lors de la transmission du message, il y a k bits déformés au plus. Dans l'espace des messages $E(n, \mathfrak{A}_2)$, délimitons un sous-ensemble $H_k \subseteq E(n, \mathfrak{A}_2)$ qui, pour tout $x, y \in H_k$, vérifie l'inégalité

$$d(x, y) > k. \tag{3-6}$$

L'ensemble H_k sera dit ensemble des mots intelligibles, le terme « mot » ayant ici la même signification que le terme « message ». Alors tout $x \notin H_k$ est un mot inintelligible. Supposons qu'à la transmission du mot $x \in H_k$ ce dernier a été mutilé et est devenu x'.

Conformément à la condition selon laquelle le nombre de déformations ne peut être supérieur à k, on a $d(x, x') \leqslant k$ et $x' \notin H_k$, c.-à-d. x' est un mot intelligible. De cette façon, la réception d'un mot inintelligible témoigne d'une déformation qui a eu lieu pendant la transmission. Les codes vérifiant la condition (3-6) sont appelés *codes à détection d'erreurs.*

Exemple 3-3. Dans l'ensemble E $(3, \mathfrak{A}_2)$, séparons l'ensemble des mots intelligibles vérifiant la condition $d(x, y) = 2$; ce sera

$$H_1 = \{000, 101, 011, 110\}.$$

La déformation d'un chiffre quelconque dans ces mots les rend inintelligibles, permettant de détecter, par là même, une erreur unique.

L'ensemble H_1 formant le code à détection d'erreur unique dans les mots de longueur n se construit de la façon suivante. Considérons l'ensemble E $(n-1, \mathfrak{A}_2)$, c.-à-d. l'ensemble des mots de longueur $n-1$. L'ensemble H_1 s'obtient en ajoutant à cet ensemble un chiffre supplémentaire choisi de façon que le nombre total d'unités dans les mots $x \in H_1$ soit pair.

Exemple 3-4. Pour $n = 4$, on a:

$$E(3, \mathfrak{A}_2) = \{000, 001, 010, 100, 011, 101, 110, 111\}.$$

Alors

$$H_1 = \{0000, 0011, 0101, 1001, 0110, 1010, 1100, 1111\}.$$

Il est commode de chercher le mot déformé en effectuant l'opération d'addition en module 2 d'après les règles

$$0 \oplus 0 = 0; \quad 0 \oplus 1 = 1; \quad 1 \oplus 0 = 1; \quad 1 \oplus 1 = 0. \tag{3-7}$$

C'est ainsi que dans le mot $x = a_1 a_2 \ldots a_n$ la quantité

$$\beta = a_1 \oplus a_2 \oplus \ldots \oplus a_n$$

sera le zéro ou l'unité suivant que le nombre de symboles du mot ayant la valeur de l'unité est pair ou impair.

Remarquons que le code établi dans l'exemple 3-4, possédant quatre chiffres et détectant l'erreur, permet de former $2^4 = 16$ mots différents au total; or, de ce nombre, il n'y a que 8 mots intelligibles, c.-à-d. convenant à la transmission par le canal de communication. Les codes dans lesquels le nombre de mots intelligibles est inférieur au nombre total de mots possibles portent le nom de *codes redondants.* La redondance est une condition nécessaire de construction des codes à détection d'erreurs.

On a pu construire aussi des codes permettant de corriger les erreurs commises. Supposons de nouveau qu'il n'y a que k bits du code qui ont été déformés au cours de la transmission. L'ensemble des mots intelligibles $H_k = E(n, \mathfrak{A}_2)$ se définit par la condition

$$d(x, y) > 2k \tag{3-8}$$

pour tout $x, y \in H_k$.

Soient deux mots quelconques $x, y \in H_k$. Supposons que le mot x a été déformé en x'. Alors $d(x, x') \leqslant k$. De l'inégalité triangulaire,

on a:

$$d(x', y) \geqslant d(x, y) - d(x, x') > 2k - k = k. \quad (3\text{-}9)$$

Par conséquent,

$$d(x, x') < d(x', y). \quad (3\text{-}10)$$

Ainsi donc, la distance entre le mot erroné x' et le mot initial x est inférieure à la distance séparant x' de tout autre mot intelligible. Trouvant le mot intelligible le plus proche de x', on reconstitue le message initial x. Les codes vérifiant la condition (3-8) s'appellent *codes à correction d'erreurs*. Les problèmes de réalisation pratique des codes à correction d'erreurs sont assez compliqués et font l'objet d'ouvrages spéciaux [19].

3-3. ESPACES LINÉAIRES NORMÉS

a) Espace linéaire

Au commencement de ce chapitre, on a défini l'espace comme un ensemble muni d'une structure. On a considéré les espaces métriques dont la structure était définie par ce qu'à chaque couple d'éléments on associait un nombre réel possédant des propriétés déterminées et appelé métrique. Or, la métrique ne peut aucunement caractériser la totalité des propriétés structurales des différents espaces. En particulier, parmi les propriétés structurales importantes d'un ensemble X des nombres réels ou complexes, on signale la possibilité d'obtenir certains éléments de l'ensemble à partir d'autres éléments en additionnant ces éléments ou en multipliant un élément par un scalaire. Les ensembles jouissant de ces propriétés se rapportent à la classe des *espaces linéaires*. Les espaces linéaires doivent satisfaire aux conditions suivantes:

1) quels que soient les éléments $x, y \in X$, il leur correspond un élément et un seul $z \in X$, appelé leur *somme* et noté $x + y$, tel que:

$x + y = y + x$ (commutativité);

$x + (y + v) = (x + y) + v$ (associativité);

il existe dans X un élément 0 tel que $x + 0 = x$ quel que soit $x \in X$ (existence d'un élément nul);

quel que soit $x \in X$, il existe dans X un élément $-x$ tel que $x + (-x) = 0$ (existence d'un élément opposé);

2) quels que soient le nombre α et l'élément $x \in X$, il leur correspond un élément et un seul $\alpha x \in X$ tel que

$$(\alpha + \beta) x = \alpha x + \beta x;$$

$$\alpha (x + y) = \alpha x + \alpha y.$$

Les conditions 1) et 2) portent le nom de conditions d'additivité et d'homogénéité de l'espace linéaire.

Donnons quelques exemples d'espaces linéaires:

1. Ensemble des nombres réels R avec les opérations d'addition et de multiplication définies de façon habituelle.

2. Ensemble R_n de tous les n-uplets ordonnés des nombres réels sur lequel les opérations d'addition et de multiplication par un nombre sont définies comme suit. Si $x, y \in R^n$ et $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, alors $x + y = (x_1 + y_1, \ldots, x_n + y_n)$; $\alpha x = (\alpha x_1, \ldots, \alpha x_n)$.

Pour $n = 2$ et $n = 3$, ces opérations coïncident avec les règles habituelles d'opérations sur des vecteurs. Le vecteur nul est le vecteur $(0, \ldots, 0)$ dont toutes les composantes sont nulles.

3. Espace des fonctions $C_{[a, b]}$ dans lequel, quels que soient $x(t), y(t) \in C_{[a, b]}$, on entend par somme $x(t) + y(t)$ la somme des valeurs de $x(t)$ et de $y(t)$ prises pour les mêmes valeurs de t, et par fonction $\alpha x(t)$, une fonction nouvelle résultant de $(x)\,t$ par multiplication de toutes ses valeurs par α. La fonction nulle est la fonction $f(t) \equiv 0$ s'annulant identiquement sur tout l'intervalle $[a, b]$.

b) Espace linéaire normé

Un espace linéaire ne se trouve décrit de façon exhaustive qu'au moment où les propriétés d'additivité et d'homogénéité se complètent par la possibilité de mesurer les valeurs des éléments eux-mêmes. C'est ainsi qu'on ne peut comparer les vecteurs tant qu'on n'a pas convenu de la signification de la valeur (longueur) d'un vecteur. Introduisant dans l'espace linéaire des appréciations numériques de la valeur des éléments isolés, on aboutit à la notion d'*espace linéaire normé*, appelé parfois *espace de Banach*.

Un espace linéaire est dit normé si pour tout $x \in X$ il existe un nombre non négatif $\| x \|$, appelé *norme de* x et qui satisfait aux conditions suivantes:

$\| x \| = 0$ si et seulement si $x = 0$;

$\| \alpha x \| = | \alpha | \cdot \| x \|$;

$\| x + y \| \leqslant \| x \| + \| y \|$ (inégalité triangulaire).

On s'assure facilement que la quantité $\| x - y \|$ possède toutes les propriétés de la distance $d(x, y)$ dans l'espace métrique. En effet,

$\| x - y \| = 0$ si $x - y = 0$, c.-à-d. si $x = y$;

se rappelant que $y - x = -(x - y)$, on trouve:

$$\| y - x \| = | -1 | \cdot \| x - y \| = \| x - y \|;$$

$$\| x - y \| = \| (x - z) + (z - y) \| \leqslant \| x - z \| + \| z - y \|.$$

Ainsi donc, l'espace linéaire normé est un espace métrique muni d'une métrique

$$d(x, y) = \| x - y \|. \tag{3-11}$$

Complétés des propriétés d'additivité et d'homogénéité, tous les espaces métriques considérés plus haut se transforment en espaces linéaires normés. Pour ces derniers il existe des notations spéciales, à savoir:

1) espace $C_2^{(n)}$ ou E_n de norme

$$\| x \| = \{ \sum_{i=1}^{n} | x_i |^2 \}^{1/2} \quad \text{ou} \quad \| x \| = | x | \quad \text{pour} \quad n = 1 ; \qquad (3\text{-}12)$$

2) espace $C_1^{(n)}$ de norme

$$\| x \| = \sum_{i=1}^{n} | x_i | ; \qquad (3\text{-}13)$$

3) espace $C^{(n)}$ de norme

$$\| x \| = \max \{ | x_1 |, \ldots, | x_n | \} ; \qquad (3\text{-}14)$$

4) espace $C_{[a, b]}$ des fonctions continues sur l'intervalle $[a, b]$, de norme

$$\| f \| = \max_{a \leq t \leq b} | f(t) |. \qquad (3\text{-}15)$$

3-4. UTILISATION DES ESPACES MULTIDIMENSIONNELS DANS CERTAINS PROBLÈMES DE CYBERNÉTIQUE

a) Lissage des erreurs des données expérimentales

Le résultat de l'observation d'une grandeur physique y représente le plus souvent une suite de valeurs mesurées de cette grandeur $(x_1, \ldots, x_n) = x$, les résultats de mesure étant généralement entachées d'erreurs dues à l'imperfection de l'expérience et à l'influence de facteurs étrangers différents. La grandeur y elle-même peut ne pas demeurer constante mais varier selon une certaine loi. La tâche de l'expérience est la détermination d'une vraie valeur de la grandeur en question.

La situation décrite peut être représentée en termes de la théorie de la transmission des messages si y est considéré comme un certain message relatif à la valeur de la grandeur donnée, message qui a subi des déformations par suite de l'imperfection de l'expérience. Considérant le message transmis y et la valeur reçue x comme des points de l'espace des messages, on peut apprécier le degré de déformation du message reçu d'après la valeur de la distance $d(x, y)$, dont le mode de détermination est fonction de la nature de l'expérience. Dans la plupart des cas, on emploie la distance de la forme

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}, \qquad (3\text{-}16)$$

pour laquelle l'espace des messages se transforme en espace $C_2^{(n)}$.

En général, on arrive à établir, sur la base des considérations théoriques, l'ensemble Y des messages théoriquement possibles. On prend en qualité de message correct un $y \in Y$ tel qu'il soit le plus proche du message reçu x, c.-à-d. pour lequel

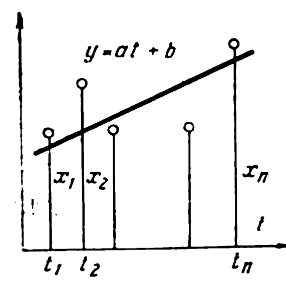


Fig. 3-3. Détermination des paramètres d'un signal linéairement variable

$$d\,(x,\ y) = \min \tag{3-17}$$

ou, ce qui est parfois plus commode,

$$d^2 = (x,\ y) = \min. \tag{3-18}$$

Le principe de recherche du message initial y par la formule (3-18) pour une distance de la forme (3-16) connaît un large emploi sous la dénomination de méthode des moindres carrés [20].

En qualité d'exemple, prenons un cas extrêmement important pour la pratique où la grandeur y varie suivant la loi linéaire

$$y = at + b. \tag{3-19}$$

Aux instants $t_1, \ldots, t_n$ on mesure les valeurs $y_k = at_k + b$, $k = 1, \ldots, n$. On voit sur la figure 3-3 les résultats de mesures $x_1, \ldots, x_n$.

Compte tenu de (3-19), mettons la condition (3-18) sous la forme

$$d^2(x,\ y) = F\,(a,\ b) = \sum_{k=1}^{n} (x_k - at_k - b)^2 = \min. \tag{3-20}$$

Dans cette expression les inconnues sont les paramètres a et b de la loi linéaire de variation de y, qu'on est en train de chercher. Egalant à zéro les dérivées partielles par rapport à a et b de la fonction $F\,(a,\ b)$, on obtient deux équations renfermant les deux inconnues; après la simplification, ces deux équations deviennent

$$\left.\begin{aligned} a \sum_{k=1}^{n} t_k^2 + b \sum_{k=1}^{n} t_k &= \sum_{k=1}^{n} t_k x_k\,; \\ a \sum_{k=1}^{n} t_k + nb &= \sum_{k=1}^{n} x_k. \end{aligned}\right\} \tag{3-21}$$

De (3-21) on tire sans peine les valeurs de a et b qui nous intéressent.

Un autre cas bien important a lieu lorsque la grandeur y demeure inchangée, c.-à-d.

$$y = c = \text{const} \tag{3-22}$$

et qu'on a fait n mesures indépendantes $x_1, \ldots, x_n$ pour l'apprécier. Compte tenu de (3-16), la condition (3-18) s'écrira

$$d^2(x, y) = F(c) = \sum_{k=1}^{n} (x_k - c)^2 = \min. \tag{3-23}$$

Dérivant cette expression par rapport à c et égalant à zéro, on trouve:

$$c = \frac{1}{n} \sum_{k=1}^{n} x_k. \tag{3-24}$$

La valeur de c trouvée à l'aide de la formule (3-24) est la *moyenne arithmétique* des valeurs $x_1, \ldots, x_n$.

Si la distance est déterminée d'une façon différente, on aboutit à des formules différentes de la moyenne. Pour plus de commodité, admettons que les valeurs $x_1, \ldots, x_n$ sont disposées dans l'ordre de croissance. Nous laissons au lecteur le soin de montrer qu'en déterminant la distance d'après la formule (3-3), la valeur moyenne

$$c = \begin{cases} x_{\frac{n+1}{2}} \text{ quand } n \text{ est impair;} \\ \text{toute valeur de } x_{\frac{n}{2}-1} \text{ à } x_{\frac{n}{2}+1} \text{ quand } n \text{ est pair.} \end{cases} \tag{3-25}$$

Parfois, pour fixer les idées, on pose pour n pair

$$c = \frac{1}{2}\left(x_{\frac{n}{2}-1} + x_{\frac{n}{2}+1}\right). \tag{3-26}$$

Quand on détermine la distance d'après la formule (3-4), la formule de la moyenne se présente sous la forme:

$$c = \frac{x_1 + x_n}{2}. \tag{3-27}$$

b) Problème d'identification des images

Le domaine de la cybernétique consacré à l'identification des images se donne pour but l'établissement du modèle d'une des propriétés fondamentales du cerveau humain: celle d'identifier objets, phénomènes et situations, grâce à laquelle l'homme s'oriente avec facilité dans les circonstances les plus compliquées [21, 22].

La théorie de l'identification est basée principalement sur les notions de *classe* et d'*image*. Les différents objets ou phénomènes de la réalité se distinguent les uns des autres d'après leurs propriétés; ils possèdent, en revanche, aussi bien des propriétés communes permettant de grouper les objets en ensembles, ou classes. C'est ainsi que les voitures de tourisme, de sport, de course constituent la

classe des voitures. Les classes peuvent être plus ou moins générales suivant les propriétés qui les caractérisent. Par exemple, la classe des voitures est un élément d'une classe plus vaste des véhicules. Ainsi donc, on entend par classe un ensemble d'objets ou de phénomènes caractérisés par des propriétés communes.

Dans chaque cas concret, on a affaire à une collection finie de classes exprimées par un ensemble fini

$$W = \{A, B, C, \ldots, P\}.$$

Toute classe contient un ensemble d'objets dont le nombre peut être aussi élevé que l'on veut et dont les propriétés sont très diverses. Or, en pratique, on n'arrive à prendre en considération qu'un nombre restreint, et souvent très faible, des différentes propriétés. Donnons à la collection des propriétés d'un objet en nombre fini l'appellation d'*image* de l'objet et représentons-la sous forme d'un vecteur à n dimensions $x = (x_1, \ldots, x_n)$ dont les composantes constituent des caractéristiques quantitatives de l'image. Un exemple d'image est donné par la représentation d'une photographie à l'écran d'un téléviseur divisée en n cellules indépendantes. Les brillances de ces cellules constituent ensemble un point dans l'espace à n dimensions, point qui est justement l'image de la photographie.

A condition que la collection choisie d'indices caractérise de façon suffisamment complète les propriétés des objets, on peut s'attendre à ce que la totalité des points correspondant aux différentes images d'une même classe occupe dans l'espace des images un domaine distinct des domaines correspondant aux images d'autres classes.

On considère que le problème d'identification des images est résolu en principe chaque fois qu'on a pu tracer dans l'espace des images les surfaces séparant cet espace en domaines dont les points appartiennent aux images de classes différentes. Une des méthodes possibles de résolution de ce problème consiste dans l'étude préalable des propriétés des échantillons isolés d'objets de chaque classe.

Supposons qu'il n'y a que deux classes d'objets A et B. Désignons par A_0 et B_0 les ensembles des échantillons préalablement étudiés faisant partie de ces classes. Placés dans l'espace des images, ces échantillons se présentent, pour le cas de deux dimensions, comme il est montré sur la figure 3-4,*a*, où les points correspondant aux échantillons de A_0 sont marqués par des ronds, et ceux de B_0, par des triangles.

Soit x une image d'un nouvel objet à étudier. Prenons pour mesure de parenté de cette image envers les classes A et B les quantités $S(x, A)$ et $S(x, B)$ qui représentent, par exemple, les carrés moyens des distances séparant le point x des points correspondant aux différents échantillons des classes A et B. Etant donné que $A_0 = \{a_1, \ldots, a_m\}$ et $B_0 = \{b_1 \ldots, b_l\}$, la formule (3-24) de la

moyenne nous donne:

$$\left.\begin{aligned} S(x, A) &= \frac{1}{m} \sum_{i=1}^{m} d^2(x, a_i); \\ S(x, B) &= \frac{1}{l} \sum_{j=1}^{l} d^2(x, b_j). \end{aligned}\right\} \quad (3\text{-}28)$$

Dans l'espace des images, les images des deux classes seront séparées par une surface C dont les points satisfont à la condition

$$S(x, A) = S(x, B). \quad (3\text{-}29)$$

Dans certains cas, la surface C est telle qu'une image sera identifiée avec erreur comme appartenant à une classe étrangère (voir fig. 3-4,*a*). Comme critère de qualité d'identification, on adopte le nombre relatif d'images incorrectement identifiées.

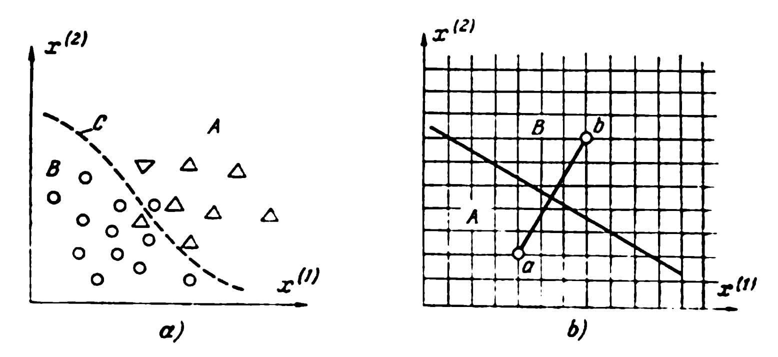


Fig. 3-4. Division en classes de l'espace des échantillons

Exemple 3-5. Dans chacune des classes A et B on a pris un échantillon $a_1 \in A$ et $b_1 \in B$ de telle façon qu'ils puissent exprimer dans l'espace bidimensionnel les propriétés les plus caractéristiques de ces classes. On demande de tracer la ligne partageant l'espace des images en deux domaines correspondant aux classes A et B.

Proposons-nous de définir la distance d'après la formule (3-2) par la longueur du segment reliant deux points dans l'espace des images. Conformément à (3-29), la ligne de partage doit avoir tous ses points à la même distance de a_1 et de b_1. Ce se rala perpendiculaire dressée au milieu du segment reliant les points a_1 et b_1 (fig. 3-4,*b*).

3-5. IMAGES GÉOMÉTRIQUES DANS L'ESPACE MULTIDIMENSIONNEL

a) Notion d'hypersphère

Soit X un ensemble de points dans l'espace multidimensionnel. On entend par *hypersphère* une surface fermée dont tous les points se trouvent à la même distance r d'un point fixé a. Donc, c'est un

ensemble $X(a, r) \subseteq E_n$ défini comme suit:

$$X(a, r) = \{x \in E_n : d(a, x) = r\}. \quad (3\text{-}30)$$

Le point fixé a est le *centre*, et r le *rayon* de l'hypersphère. L'ensemble

$$B(a, r) = \{x \in E_n : d(a, x) < r\} \quad (3\text{-}31)$$

est la partie intérieure de l'hypersphère, ou *boule ouverte*. L'ensemble

$$B[a, r] = \{x \in E_n : d(a, x) \leqslant r\} \quad (3\text{-}32)$$

est appelé *boule fermée*.

Une boule ouverte de rayon ε centrée en x porte le nom de *voisinage* de x et se note $V_\varepsilon(x)$.

Exemple 3-6. Dans l'espace $E_2 = \{x_1, x_2\}$, l'équation de l'hypersphère peut s'écrire $d^2(a, x) = r^2$. La distance étant définie d'après la formule (3-7), l'hypersphère se transforme en une circonférence: $(x_1 - a_1)^2 + (x_2 - a_2)^2 = r^2$.

b) Ensembles bornés et finis

L'ensemble E_n contenant tous les points innombrables de l'espace à n dimensions sera appelé ensemble universel. Désignons par X un sous-ensemble de E_n. On dit que X est *borné* si la distance séparant deux points arbitraires de cet ensemble est bornée, c.-à-d. s'il y a un nombre M tel que pour tout $x^{(1)}, x^{(2)} \in X$ on ait

$$d(x^{(1)}, x^{(2)}) \leqslant M. \quad (3\text{-}33)$$

Pour que l'ensemble X soit borné, il faut et il suffit qu'il se trouve dans une hypersphère, c.-à-d. qu'il existe un point a et un nombre r tels que pour tout $x \in X$ on ait

$$d(x, a) \leqslant r. \quad (3\text{-}34)$$

L'ensemble X est *fini* s'il contient un nombre fini de points. Un ensemble fini est toujours borné.

Exemple 3-7. Dans l'espace E_2 l'ensemble $X = \{(x_1, x_2) \in E_2 : x_1^2 + x_2^2 < 1\}$ est un ensemble borné infini.

Exemple 3-8. Dans l'espace E_2 l'ensemble X constitué par cinq points $X = \{(0,0), (0,1), (1,0), (1,1), (1/2, 1/2)\}$ est un ensemble fini.

c) Ensembles ouverts et fermés

Un point x est dit *intérieur* à l'ensemble X s'il y a un voisinage $V_\varepsilon(x)$ dont tous les points appartiennent à X. L'ensemble X dont tous les points sont intérieurs est dit *ensemble ouvert*.

Exemple 3-9. L'intervalle $(a, b) = \{x \in R : a < x < b\}$, où R est un ensemble des nombres réels, est un ensemble ouvert. En effet, pour tout $x \in (a, b)$ le voisinage $V_\varepsilon(x)$, où $\varepsilon = \min(x - a, b - x)$, est contenu dans (a, b).

Exemple 3-10. La boule ouverte $B(a, r)$ est un ensemble ouvert. En effet, si $x \in B(a, r)$, alors $d(a, x) < r$. Posons $\varepsilon = r - d(a, x)$. Alors $V_\varepsilon(x) = B(x, \varepsilon) \subset B(a, r)$.

Le point x est dit *point limite* de l'ensemble X si tout voisinage $V_\varepsilon(x)$ de ce point contient une infinité de points de X. Si en outre $x \in X$, c'est un point limite appartenant à X. Par exemple, tous les points intérieurs à l'ensemble X sont des points limites appartenant à X.

Si x est un point limite de l'ensemble X et que $x \notin X$, c'est un point limite n'appartenant pas à X. Ce sera par exemple le cas des points a et b de l'ensemble (a, b). Un ensemble fini est dépourvu de points limites.

Un ensemble X est dit *fermé* s'il contient tous ses points limites. Tels sont par exemple un segment quelconque $[a, b]$ de la droite numérique, une boule fermée $B[a, r]$, ainsi que tout ensemble fini.

d) Notion d'hyperplan

L'*hyperplan* est une généralisation de la notion de plan pour l'espace multidimensionnel. On montre en Géométrie analytique que toute équation linéaire par rapport aux coordonnées définit un plan. Si les coordonnées d'un point courant sont x_1, x_2, x_3, l'équation du plan dans l'espace tridimensionnel a la forme:

$$a_1x_1 + a_2x_2 + a_3x_3 = c. \tag{3-35}$$

Si l'on considère le triplet fixe de nombres (a_1, a_2, a_3) comme le vecteur a mené de l'origine des coordonnées au point de coordonnées a_1, a_2 et a_3 et $x = (x_1, x_2, x_3)$ comme le vecteur définissant la position du point courant dans l'espace E_3, le premier membre de l'équation (3-35) représentera le produit scalaire des vecteurs a et x et l'équation

$$ax = c \tag{3-36}$$

sera l'équation vectorielle du plan perpendiculaire au vecteur $a = (a_1, a_2, a_3)$.

Dans le cas de l'espace multidimensionnel E_n l'équation (3-36) définira l'hyperplan perpendiculaire au vecteur fixé $a = (a_1, \ldots, a_n)$ et dont les coordonnées courantes se définissent par le vecteur $x = (x_1, \ldots, x_n)$. Désignant l'ensemble des points de l'hyperplan par $L(x)$ et développant le produit scalaire des vecteurs a et x, on aboutit à la définition suivante de l'hyperplan:

$$L(x) = \{x \in E_n : \sum_{i=1}^{n} a_ix_i - c = 0\}. \tag{3-37}$$

Les ensembles

$$L^+ = \{x \in E_n : \sum_{i=1}^{n} a_i x_i - c > 0\} \qquad (3\text{-}38)$$

et

$$L^- = \{x \in E_n : \sum_{i=1}^{n} a_i x_i - c < 0\} \qquad (3\text{-}39)$$

sont appelés demi-espaces ouverts définis par l'hyperplan L.

e) Equation d'un segment. Moyenne pondérée des éléments d'un ensemble

Pour obtenir l'équation de la droite passant par les points $x^{(1)}$ et $x^{(2)}$ dans l'espace E_n, on prendra sur cette droite un point x de façon qu'il divise le segment $(x^{(1)}, x^{(2)})$ en rapport $w: 1 - w$. Puisque le rapport des segments de droite reste le même que le rapport de leurs projections sur les axes de coordonnées, pour les projections sur le k-ième axe on obtient la relation

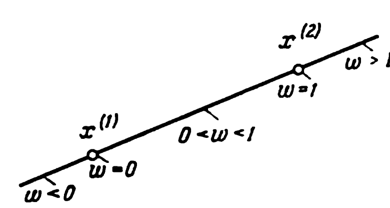


Fig. 3-5. Valeurs de w en divers points de la droite

$$\frac{x_k - x_k^{(1)}}{x_k^{(2)} - x_k} = \frac{w}{1-w} \qquad (3\text{-}40)$$

ou

$$x_k = (1-w)\, x_k^{(1)} + w x_k^{(2)}. \qquad (3\text{-}41)$$

Posant $k = 1, \ldots, n$, on obtient le système d'équations définissant la droite

$$\left.\begin{array}{l} x_1 = (1-w)\, x_1^{(1)} + w x_1^{(2)}\,; \\ \cdots\cdots\cdots\cdots \\ x_n = (1-w)\, x_n^{(1)} + w x_n^{(2)}. \end{array}\right\} \qquad (3\text{-}42)$$

Ce système d'équations peut être écrit de façon plus condensée sous forme vectorielle

$$x = (1-w)\, x^{(1)} + w x^{(2)}. \qquad (3\text{-}43)$$

L'équation (3-43) est l'équation de la droite passant par les points $x^{(1)}$ et $x^{(2)}$, et ceci de la sorte qu'aux différentes valeurs de w (fig. 3-5) correspondent les différents points de cette droite.

En particulier, pour $0 \leqslant w \leqslant 1$ cette équation définit le segment compris entre les points $x^{(1)}$ et $x^{(2)}$.

Il est plus commode d'écrire l'équation d'un segment sous la forme

$$x = w_1 x^{(1)} + w_2 x^{(2)}\,; \quad w_1,\ w_2 \geqslant 0\,; \quad w_1 + w_2 = 1. \qquad (3\text{-}44)$$

La quantité x tirée de (3-44) porte le nom de *moyenne pondérée* des éléments $x^{(1)}$ et $x^{(2)}$ de poids w_1 et w_2. Cette appellation a une signification physique simple. Si l'on place en $x^{(1)}$ et $x^{(2)}$ les poids w_1 et w_2, le point x se trouvera au centre de gravité du système.

La notion de moyenne pondérée peut être étendue à un plus grand nombre de points. Etant donné les éléments $x^{(1)}, \ldots, x^{(m)}$, la moyenne pondérée x est

$$x = \sum_{i=1}^{m} w_i x^{(i)}, \quad w_i > 0, \quad \sum_{i=1}^{m} w_i = 1. \tag{3-45}$$

3-6. ENSEMBLES CONVEXES ET LEURS PROPRIÉTÉS

a) Définition d'un ensemble convexe

Soit X un ensemble dans l'espace E_n. L'ensemble X est dit *convexe* si le segment reliant deux points arbitraires de cet ensemble est contenu dans cet ensemble. En d'autres mots, X est un ensemble convexe si pour tout $x^{(1)}, x^{(2)} \in X$ et pour tout $w_1, w_2 > 0$, $w_1 + w_2 = 1$ on a $w_1 x^{(1)} + w_2 x^{(2)} \in X$.

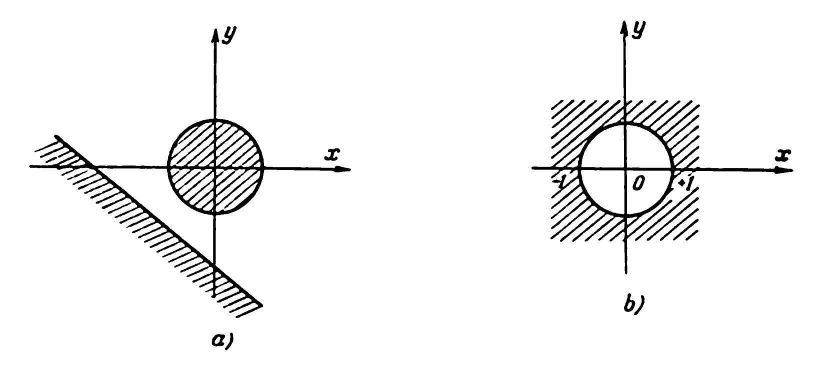


Fig. 3-6. Exemples d'ensembles convexes (a) et non convexes (b)

Les exemples d'ensembles convexes dans l'espace $E_2 = \{(x, y)\}$ sont: $x^2 + y^2 < 1$; $x^2 + y^2 \leqslant 1$; tout le plan E_2; les demi-plans $ax + by - c > 0$ et $ax + by - c < 0$ (fig. 3-6,a). D'autre part, aucun des deux ensembles définis par les équations $x^2 + y^2 \geqslant 1$, $x^2 + y^2 = 1$ (fig. 3-6,b) n'est convexe. En effet, le point (0, 0) n'appartenant pas à ces ensembles appartient au segment reliant les points (1, 0) et (—1,0) des ces ensembles.

Théorème 3-1. *L'intersection d'ensembles convexes est un ensemble convexe.*

Démonstration. Considérons l'intersection $X \cap Y$ d'ensembles convexes X et Y. Soient x et y deux points quelconques de

$X \cap Y$. Ils appartiennent donc tant à X qu'à Y. Les ensembles X et Y étant convexes, le segment joignant les points x et y appartient entièrement à X et à Y, donc à l'intersection $X \cap Y$. Par conséquent, $X \cap Y$ est un ensemble convexe.

b) Enveloppe convexe d'un ensemble fini

Soit $A = \{a_1, \ldots, a_m\}$ un ensemble fini de points dans l'espace E_n. Un ensemble fini n'est pas convexe. Or, les points de l'ensemble fini A peuvent être les éléments de quelques ensembles convexes, par exemple, de S_1, S_2, etc., comme il est montré à la figure 3-7. Dans ce cas A est un sous-ensemble des ensembles convexes $S_1, S_2, \ldots$

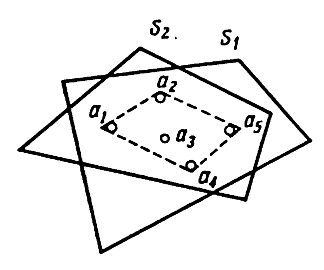


Fig. 3-7. Sur la recherche de l'enveloppe convexe d'un ensemble fini

On entend par *enveloppe convexe* co (A) d'un ensemble A l'intersection de tous les ensembles convexes dont A est un sous-ensemble. En particulier, si $A \subset S_1$ et $A \subset S_2$, alors co $(A) \subseteq S_1 \cap S_2$.

Il découle de cette définition que l'enveloppe convexe co (A) est le plus petit ensemble convexe contenant A. En effet, s'il existait un ensemble convexe S tel que $A \subset S$ et $S \subseteq$ co (A), on aurait co $(A) \subseteq S \cap$ co (A), c.-à-d. co $(A) \subseteq S$. Il en résulte que $S =$ co (A).

Théorème 3-2. *L'enveloppe convexe d'un ensemble fini A est l'ensemble des moyennes pondérées des éléments de A.*

Soient $A = \{a_1, \ldots, a_m\}$ un ensemble fini et w_i, $i = 1, \ldots, m$ des nombres réels tels que

$$w_i \geqslant 0; \quad \sum_{i=1}^{m} w_i = 1. \tag{3-46}$$

La quantité

$$x = \sum_{i=1}^{m} w_i a_i \tag{3-47}$$

s'appelle moyenne pondérée des éléments de A. L'ensemble S qui a pour éléments les quantités x, définies par la formule (3-47) pour tout w_i dans les limites définies par (3-46), représente l'ensemble des moyennes pondérées des éléments de l'ensemble A. Il s'agit de démontrer que

$$S = \text{co}\,(A). \tag{3-48}$$

D é m o n s t r a t i o n. Pour que l'ensemble S soit l'enveloppe convexe de l'ensemble A, il faut d'abord qu'il soit convexe et, ensuite, qu'il soit le plus petit ensemble convexe contenant A. Montrons que ces deux conditions sont vérifiées.

1. Prenons deux systèmes arbitraires de poids w_i et w'_i définissant deux points de S:

$$x = \sum_{i=1}^{m} w_i a_i, \quad x' = \sum_{i=1}^{m} w'_i a_i. \tag{3-49}$$

Choisissons un point arbitraire x'' sur le segment reliant x et x':

$$x'' = (1-w)\,x + wx' = \sum_{i=1}^{m} [(1-w)\,w_i + ww'_i]\,a_i = \sum_{i=1}^{m} w''_i a_i, \tag{3-50}$$

où

$$w''_i = (1-w)\,w_i + ww'_i. \tag{3-51}$$

La quantité w''_i est non négative, car elle représente la moyenne pondérée de deux nombres non négatifs w_i et w'_i. En outre,

$$\sum_{i=1}^{m} w''_i = (1-w)\sum_{i=1}^{m} w_i + w\sum_{i=1}^{m} w'_i = 1. \tag{3-52}$$

Donc, x'' est la moyenne pondérée des éléments de l'ensemble A, c.-à-d. que $x'' \in S$. Par conséquent, S est un ensemble convexe.

2. Soit T un ensemble convexe quelconque contenant A. Montrons qu'il contient aussi S.

Par hypothèse, T contient tous les éléments a_i. Ensuite, l'élément

$$x_2 = \frac{w_1}{w_1+w_2}\,a_1 + \frac{w_2}{w_1+w_2}\,a_2 \tag{3-53}$$

est la moyenne pondérée de a_1, $a_2 \in T$, de sorte que $x_2 \in T$. D'une manière analogue il vient que

$$x_3 = \frac{w_1+w_2}{w_1+w_2+w_3}\,x_2 + \frac{w_3}{w_1+w_2+w_3}\,a_3 = \\ = \frac{w_1a_1}{w_1+w_2+w_3} + \frac{w_2a_2}{w_1+w_2+w_3} + \frac{w_3a_3}{w_1+w_2+w_3} \tag{3-54}$$

est la moyenne pondérée de x_2, $a_3 \in T$, de sorte que $x_3 \in T$. Poursuivons l'examen de cette suite de points; il vient que le point

$$x_m = \frac{w_1+\ldots+w_{m-1}}{w_1+\ldots+w_m}\,x_{m-1} + \frac{w_m}{w_1+\ldots+w_m}\,a_m = \sum_{i=1}^{m} w_i a_i \in S \tag{3-55}$$

appartient aussi à T. Ainsi donc, tout $x \in S$ appartient aussi à T, c.-à-d. $S \subseteq T$. Or, T est un ensemble convexe arbitraire contenant A. Il s'ensuit que S est le plus petit ensemble convexe contenant A.

Le théorème démontré permet d'établir la forme de l'enveloppe convexe d'un ensemble fini. Prenons le cas bidimensionnel. A la

figure 3-8 est représenté un ensemble fini $A = \{a_1, a_2, a_3, a_4, a_5\}$; on y voit aussi un polygone convexe dont les sommets sont les éléments de A. Il est facile de voir que tout point situé au sommet, sur un côté ou à l'intérieur du polygone convexe peut être représenté comme la moyenne pondérée d'au moins trois de ses sommets. Par exemple, x_0 = moy.pond. (a_1, a_2), x_1 = moy.pond. (x_0, a_3) = = moy. pond. (a_1, a_2, a_3). Par contre, un point x_2 extérieur au polygone ne peut être la moyenne pondérée d'aucun couple de points intérieurs, donc ne peut non plus être exprimé comme la moyenne

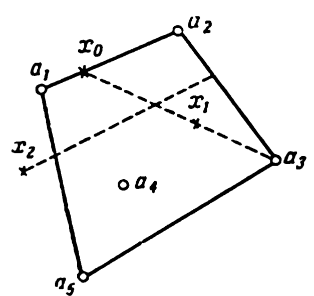


Fig. 3-8. Enveloppe convexe d'un ensemble fini

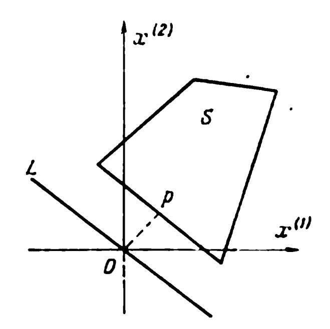


Fig. 3-9. Sur la construction de l'hyperplan d'appui

pondérée des sommets du polygone. Ainsi donc, l'enveloppe convexe d'un ensemble fini A sur le plan est un polygone convexe dont les sommets sont les éléments de A.

Cette conclusion s'étend sans peine au cas d'un espace arbitraire E_n. L'enveloppe convexe d'un ensemble fini A dans l'espace E_n est un polyèdre convexe dont les sommets sont les éléments de A. Tout point intérieur ou frontière d'un tel polyèdre peut être représenté comme la moyenne pondérée de $(n - 1)$ sommets au maximum.

En terminant, proposons sans démonstration deux théorèmes qui seront utilisés pour argumenter certaines propositions de la programmation linéaire et de la théorie des jeux.

Théorème 3-2. (théorème de l'hyperplan d'appui). *Soient S un ensemble convexe et O un point arbitraire n'appartenant pas à S. Il doit exister alors un hyperplan L passant par O de telle façon que S se trouve dans un seul des demi-espaces engendrés par L* (ce fait est intuitivement clair pour les cas de l'espace bi et tridimensionnel, voir fig. 3-9).

Commençons à déplacer l'hyperplan dans l'espace en le rapprochant de S jusqu'au moment où L et S acquièrent un point commun au minimum, à condition cependant que S soit contenu dans l'un des demi-espaces engendrés par l'hyperplan L. Cet hyperplan s'appelle *hyperplan d'appui* de l'ensemble convexe S.

Si l'ensemble convexe S est l'enveloppe convexe d'un ensemble fini A, il représente un polyèdre convexe. Dans ce cas l'hyperplan d'appui passera par le sommet, par l'arête ou par la face du polyèdre (fig. 3-10). Dans chacune de ces variantes un sommet au moins du polyèdre sera contenu dans l'hyperplan d'appui.

Théorème 3-4 (théorème de l'hyperplan de séparation). *Si S et T sont des ensembles convexes sans élément commun et que l'un au moins*

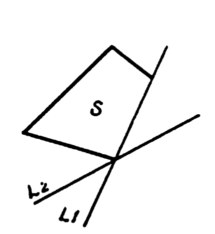


Fig. 3-10. Droites d'appui d'un ensemble convexe

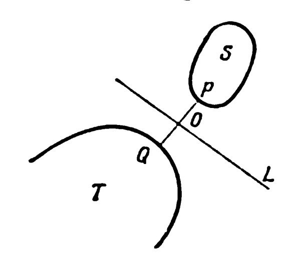


Fig. 3-11. Hyperplan de séparation

d'entre eux est borné, il existe un hyperplan tel que les ensembles S et T se situent dans des demi-espaces différents engendrés par cet hyperplan. En d'autres mots, il existe un hyperplan $px = c$, où $p = (p_1, \ldots, p_n)$ est un ensemble ordonné des nombres réels et $x = (x_1, \ldots, x_n)$ un point arbitraire, tel que $px > c$ quand $x \in S$ et $px < c$ quand $x \in T$. L'illustration géométrique de ce théorème est donnée sur la figure 3-11.

PROBLÈMES AU CHAPITRE 3

3-1. On donne dans l'espace tridimensionnel trois points $x = (1, 0, 2)$, $y = (3, 4, 0)$, $z = (1, 2, 3)$. Vérifier pour ces points l'axiome d'inégalité triangulaire dans les espaces $C_2^{(3)}$, $C_1^{(3)}$, $C^{(3)}$.

3-2. Tracer sur le plan (x, y) le cercle de rayon r et de centre a dans les espaces $C_2^{(3)}$, $C_1^{(3)}$, $C^{(3)}$.

3-3. Démontrer les formules (3-25) et (3-27) de la moyenne.

3-4. Décrire le changement qualitatif de la ligne séparant les classes A et B de l'exemple 3-5 qui se produit si l'on prend dans A un second échantillon a_2. Examiner les cas $a_2 = (2, 1)$, $a_2 = (2, 3)$, $a_2 = (6, 3)$.

3-5. De quelle forme sera la ligne séparant les classes A et B sur la figure 3-4,*b* si la distance est définie:

a) d'après la formule (3-3);

b) d'après la formule (3-4)?

3-6. Montrer que la boule ouverte $B(0, r)$ dans l'espace E_n est un ensemble convexe.

3-7. Soient $[0, 3]$, $[5, 7]$, $[0, 3] \cup [5, 7]$ des ensembles dans l'espace R des nombres réels. Lesquels de ces ensembles sont convexes?

3-8. Montrer que le point x_1 de la figure 3-8 est la moyenne pondérée des sommets a_1, a_2 et a_3.

CHAPITRE 4

ÉLÉMENTS D'ALGÈBRE DE LA LOGIQUE

4-1. OPÉRATIONS LOGIQUES

a) Notion de propositions

On a vu plus haut que la définition d'un ensemble peut être donnée par deux méthodes: par énumération et par description, c.-à-d. par indication de la propriété inhérente aux éléments de l'ensemble. La méthode descriptive de définition d'ensembles est un pont entre la théorie des ensembles et la théorie des propositions, cette dernière représentant le premier chapitre, et aussi le plus élémentaire, d'une discipline scientifique dite logique mathématique [23 à 32].

On entend par *proposition* une assertion quelconque qui peut être soit vraie, soit fausse. Appliquons le concept de propositions aux éléments d'un ensemble universel I. Les différents éléments de cet ensemble ont des propriétés différentes et peuvent donc former différents groupes qui sont sous-ensembles de I. Par exemple, si I est l'ensemble des étudiants du groupe, il peut avoir comme sous-ensembles: X l'ensemble des étudiants qui ont toutes leurs notes excellentes; Y l'ensemble des étudiants logés au foyer d'étudiants; Z l'ensemble des étudiants qui touchent la bourse, et ainsi de suite.

Une fois que l'on s'est rendu compte des propriétés caractérisant les différents sous-ensembles, on peut formuler des assertions déterminées selon lesquelles tel ou tel élément de I possède ou non la propriété en question. De pareilles assertions sont justement des propositions. Donnons-en quelques exemples: « il a toutes ses notes excellentes », « il loge au foyer », « il touche la bourse ».

Nous ferons par la suite abstraction de toutes les propriétes des propositions sauf une: toute proposition appliquée à l'élément considéré de l'ensemble universel I est soit vraie, soit fausse. C'est ainsi que la proposition « l'étudiant Ivanov a toutes ses notes excellentes » est vraie si Ivanov fait partie du sous-ensemble X et fausse dans le cas contraire.

Aussi dit-on que X est l'*ensemble de vérité* (ou *ensemble des valeurs de vérité*) de la proposition « il a toutes ses notes excellentes ». Les ensembles de vérité des propositions « il loge au foyer » et « il touche la bourse » sont respectivement les ensembles Y et Z.

L'ensemble de vérité d'une proposition peut s'avérer être vide. On dit alors que la proposition est *identiquement fausse*. Pour un

groupe d'étudiants c'est le cas de la proposition « il a plus de cinquante ans ». Il peut arriver que l'ensemble de vérité d'une proposition se confonde avec l'ensemble universel I ; c'est une proposition *identiquement vraie*, ou *tautologique*. Pour un groupe d'étudiants, la proposition identiquement vraie est « il a moins de cinquante ans ».

b) Propositions élémentaires et composées

Désignons les propositions par des lettres minuscules de l'alphabet latin et associons à chacune d'elles les valeurs numériques 1 ou 0 suivant que la proposition est vraie ou fausse. Supposons, par exemple, que x désigne la proposition « il a toutes ses notes excellentes ». Elle a comme valeurs numériques

$$x = \begin{cases} 1 \text{ si } x \text{ est vrai, c.-à-d. } x \in X; \\ 0 \text{ si } x \text{ est faux, c.-à-d. } x \notin X. \end{cases} \quad (4\text{-}1)$$

Les propositions x, y, z ayant leurs valeurs de vérité dans des ensembles élémentaires X, Y, Z sont des *propositions élémentaires*. Or, il arrive aussi que l'ensemble de vérité Q s'obtient à partir des ensembles X, Y, Z au moyen d'une opération algébrique effectuée sur ces ensembles. A cet ensemble de vérité Q correspondra alors une *proposition composée* q. Par exemple, à l'ensemble de vérité $Q = X \cap Y$, possédant tant la propriété X (toutes les notes excellentes) que Y (loge au foyer), correspond la proposition composée « il a toutes ses notes excellentes et loge au foyer ».

Dans cet exemple nous avons obtenu une proposition composée en mettant entre deux propositions la conjonction « et ». On peut aussi former des propositions composées en employant d'autres jonctions: « ou », « si ... alors », etc. On obtient une nouvelle proposition en prenant l'inverse de la proposition initiale. A chacune de ces nouvelles propositions correspondront leurs propres ensembles de vérité sur l'ensemble universel I ; donc, les propositions composées peuvent aussi être vraies ou fausses, c.-à-d. admettre les valeurs numériques 1 et 0.

Considérant les propositions comme des variables susceptibles de valeurs 1 et 0, on arrive à définir les opérations qui, effectuées sur ces propositions, font naître des propositions nouvelles. Ces opérations exprimeront justement les liaisons employées dans le langage courant, dont on a parlé plus haut.

Les opérations effectuées sur les propositions sont appelées *opérations logiques*. L'ensemble des opérations logiques étudiées dans le texte qui suit porte le nom d'*algèbre des propositions* ou d'*algèbre de Boole*.

c) Représentation des opérations logiques

Soient quelques propositions élémentaires $x_1, \ldots, x_N$ dont chacune peut être vraie ou fausse, donc prendre des valeurs numériques 1 et 0. La totalité de ces propositions peut être considérée comme un cortège $(x_1, \ldots, x_N)$.

Supposons qu'à la suite d'une opération logique effectuée sur ces propositions on a obtenu une nouvelle proposition q qui peut être vraie ou fausse elle aussi, donc peut prendre les valeurs 1 et 0. A toute combinaison de valeurs $x_1, \ldots, x_N$ correspondra alors une valeur déterminée de $q \in \{1, 0\}$, ce qui permet de considérer une opération logique comme une application f de l'ensemble des valeurs du cortège $(x_1, \ldots, x_N)$ sur l'ensemble des valeurs de q

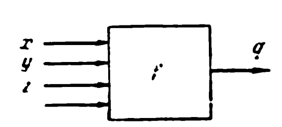


Fig. 4-1. Représentation conventionnelle d'une opération logique

$$f: (x_1, \ldots, x_N) \rightarrow \{1, 0\}.$$

Si cette application est univoque, elle définit la fonction

$$q = f(x_1, \ldots, x_N) \qquad (4\text{-}2)$$

dite *fonction booléenne*. Il découle de ce qui précède que les arguments des fonctions booléennes, aussi bien que les fonctions elles-mêmes, ne peuvent prendre que deux valeurs différentes, à savoir 1 et 0.

On emploie mois méthodes de représentation des fonctions booléennes:

1. Une formule indiquant explicitement la suite des opérations logiques effectuées sur les propositions $x_1, \ldots, x_N$ et ayant la forme de la relation (4-2).

2. Un tableau donnant les valeurs de vérité de la proposition composée q en fonction des valeurs de vérité des propositions initiales. Dans la partie gauche du tableau sont énumérées toutes les com-

Tableau 4-1

Fonctions de deux variables logiques

$x\ y$	0	x	y	xy	$x+y$	$x \oplus y$	$x \rightarrow y$	$y \rightarrow x$	1	$\bar{x}$	$\bar{y}$	$\overline{xy}$	$\overline{x+y}$	$\overline{x \oplus y}$	$\overline{x \rightarrow y}$	$\overline{y \rightarrow x}$
0 0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0
0 1	0	0	1	0	1	1	1	0	1	1	0	1	0	0	0	1
1 0	0	1	0	0	1	1	0	1	1	0	1	1	0	0	1	0
1 1	0	1	1	1	1	0	1	1	1	0	0	0	0	1	0	0

binaisons possibles des valeurs de vérité des propositions initiales $x_1, \ldots, x_N$, et dans la partie droite figurent les valeurs de vérité de la proposition composée q. Si l'on a N propositions initiales, le nombre de lignes du tableau est 2^N. On donne deux exemple de tableaux de cette espèce (tableaux 4-1 et 4-2).

Tableau 4-2

$q = \bar{x}y + z$

x	y	z	q
0	0	0	0
1	0	0	0
0	1	0	1
0	0	1	1
1	1	0	0
1	0	1	1
0	1	1	1
1	1	1	1

3. Un circuit logique, qui est une représentation graphique conventionnelle de l'opération logique (voir fig. 4-1).

En automatique et en technique de calculs, on a l'habitude de représenter les propositions sous la forme de signaux à deux niveaux ou de dispositifs bistables (relais, basculeur, tube électronique, transistor, etc.). On associe à ces deux niveaux de signaux ou deux états stables du dispositif les valeurs numériques 1 et 0 définissant les valeurs de vérité des propositions correspondantes. Le circuit logique représente alors un convertisseur de signaux qui peut être utilisé pour la commande de divers processus.

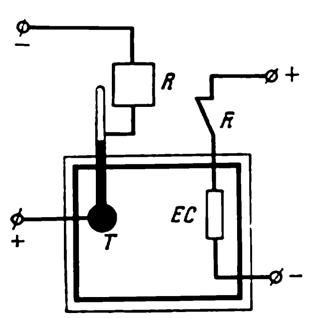


Fig. 4-2. Régulation de la température dans un thermostat

Exemple 4-1. Dans un thermostat, la température Θ est maintenue à un niveau désiré Θ_0 au moyen d'un élément chauffant *EC* (fig. 4-2) qui a deux régimes de fonctionnement: « marche » et « arrêt ». La commande de la mise en marche de l'élément chauffant s'effectue suivant la règle:

mettre en marche *EC* si $\Theta < \Theta_0$;

arrêter *EC* si $\Theta > \Theta_0$.

Considérons la proposition « la température dépasse Θ_0 » comme proposition initiale x et la proposition « mettre en marche *EC* » comme proposition composée q. La structure logique du circuit de commande de l'élément chauffant se présente alors comme suit:

$$q = \begin{cases} 1 \text{ si } x = 0; \\ 0 \text{ si } x = 1. \end{cases}$$

On voit sur la figure 4-2 une réalisation physique de cette loi de commande. Le signal x s'exprime par la longueur de la colonne de mercure d'un thermomètre T qui, pour $\Theta \geqslant \Theta_0$, ferme le contact d'un relais R et, par là même, met en marche l'élément chauffant.

4-2. ALGÈBRE DES PROPOSITIONS

a) Négation

Soit x une proposition à valeurs de vérité dans X. On désigne par $\bar{x}$ (et lit « non x ») une nouvelle proposition ayant ses valeurs de vérité dans l'ensemble $\bar{X}$ et appelée *négation* ou *inversion* de X.

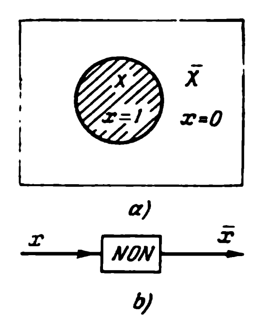


Fig. 4-3. Diagramme d'Euler-Venn et représentation conventionnelle de l'opération d'inversion

La relation entre x et $\bar{x}$ se définit au moyen du diagramme d'Euler-Venn donné sur la figure 4-3,*a*. Par définition de la négation, on a $\bar{x} = 1$ dans le domaine $\bar{X}$ où $x = 0$, et $\bar{x}=0$ dans le domaine X où $x = 1$. Cette relation s'écrit sous la forme des règles d'inversion suivantes:

$$\bar{1}=0, \quad \bar{0}=1. \tag{4-3}$$

On voit sur la figure 4-3,*b* la représentation de l'opération de négation sous la forme d'un circuit logique appelé *inverseur*.

b) Addition logique

Soient x et y deux propositions ayant leurs valeurs de vérité dans les ensembles X et Y respectivement. On désigne par $x + y$ (on écrit parfois $x \vee y$ et on lit « x ou y ») une nouvelle proposition qui a ses valeurs de vérité dans l'ensemble $X \cup Y$ et porte le nom de *somme logique* ou *disjonction* de x et y.

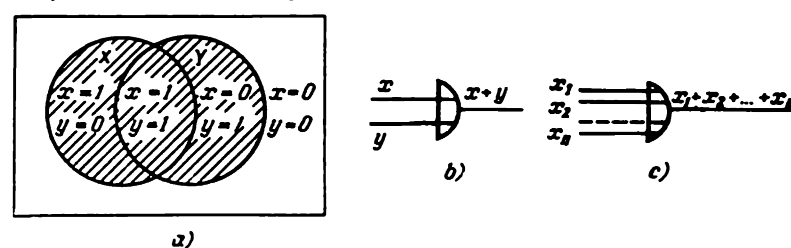


Fig. 4-4. Diagramme d'Euler-Venn et représentation conventionnelle de l'opération d'addition logique

Les relations entre les valeurs numériques de x, y et $x + y$ s'obtiennent du diagramme d'Euler-Venn de la figure 4-4,*a*. L'ensemble de vérité de la proposition $x + y$ est le domaine hachuré dans lequel

$$x+y=\begin{cases} 0 \text{ si } x=0 \text{ et } y=0; \\ 1 \text{ dans les autres cas.} \end{cases} \tag{4-4}$$

Il est commode d'écrire les relations (4-4) sous la forme de règles qui rappellent un peu les règles canoniques d'addition arithmétique:

$$\left.\begin{array}{ll} 0+0=0; & 1+0=1; \\ 0+1=1; & 1+1=1. \end{array}\right\} \qquad (4\text{-}5)$$

Le circuit logique utilisé pour représenter l'opération d'addition logique et appelé *circuit de réunion* est donné sur la figure 4-4,*b*.

La règle d'addition logique s'étend sans difficulté au cas de trois propositions et plus. Dans le cas général, on définit la somme logique de n propositions par la règle

$$x_1+x_2+\ldots+x_n=\begin{cases} 0 \text{ si } x_1=x_2=\ldots=x_n=0; \\ 1 \text{ dans les autres cas.} \end{cases} \qquad (4\text{-}6)$$

On voit le circuit logique pour ce cas sur la figure 4-4,*c*.

c) Multiplication logique

Soient x et y deux propositions à valeurs de vérité dans deux ensembles X et Y respectivement. On désigne par xy (on écrit parfois $x \wedge y$ et on lit « x et y ») une nouvelle proposition qui a ses

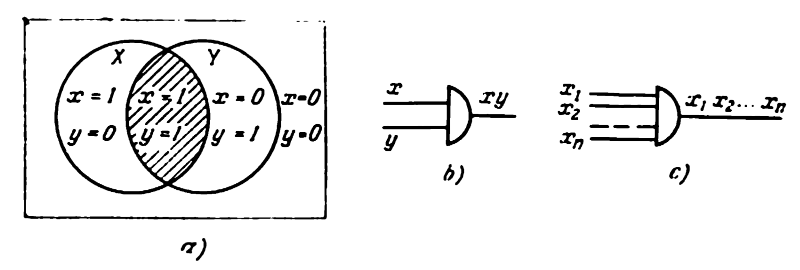


Fig. 4.5. Diagramme d'Euler-Venn et représentation conventionnelle de l'opération de multiplication logique

valeurs de vérité dans l'ensemble $X \cap Y$ et qui s'appelle *produit logique* ou *conjonction* de x et y.

L'ensemble de vérité de la proposition xy est représenté sur le diagramme d'Euler-Venn (fig. 4-5,*a*) par le domaine hachuré dans lequel

$$xy=\begin{cases} 1 \text{ si } x=1 \text{ et } y=1; \\ 0 \text{ dans les autres cas.} \end{cases} \qquad (4\text{-}7)$$

Cette relation peut être représentée sous la forme des règles de multiplication logique qui coïncident avec les règles de multiplication arithmétique:

$$0\cdot 0=0,\ 0\cdot 1=0,\ 1\cdot 0=0,\ 1\cdot 1=1. \qquad (4\text{-}8)$$

On voit sur la figure 4-5,*b* le circuit logique employé pour représenter l'opération de multiplication logique et qui s'appelle *circuit d'intersection* ou *de coïncidence*.

La règle de multiplication logique s'étend facilement au cas de trois propositions et plus. Le produit logique de *n* propositions se définit par la règle

$$x_1 x_2 \ldots x_n = \begin{cases} 1 \text{ si } x_1 = x_2 = \ldots = x_n = 1\,; \\ 0 \text{ dans les autres cas.} \end{cases} \qquad (4\text{-}9)$$

Le circuit logique traduisant ce cas est donné sur la figure 4-5,*c*.

Une des variantes du circuit de multiplication logique est l'opération du produit logique d'une proposition x par l'inversion de l'autre $\bar{y}$, qui s'écrit

$$x\bar{y} = \begin{cases} x \text{ si } y = 0\,; \\ 0 \text{ si } y = 1. \end{cases} \qquad (4\text{-}10)$$

Fig. 4-6. Circuit d'inhibition

Cette opération étant assez fréquente, on lui réserve une représentation spéciale sous la forme du circuit logique de la figure 4-6. Il est analogue au circuit de coïncidence, à la différence près que l'entrée d'inversion y ne se termine pas par une pointe mais par un rond. Comme on peut le voir de (4-10), l'amenée du signal $y = 1$ à cette entrée interdit le passage du signal x à la sortie du circuit; aussi l'entrée y porte-t-elle le nom d'entrée *inhibitive*, tandis que le circuit de la figure 4-6 est appelé *circuit d'inhibition*. L'opération logique elle-même correspondant aux relations (4-10) s'appelle opération d'*inhibition*.

d) Fonctions booléennes

Les propositions isolées auxquelles on a assigné les valeurs de vérité 1 et 0 peuvent être considérées comme des variables binaires; on les appellera par la suite *variables logiques*. Les opérations de négation, d'addition logique et de multiplication logique sont des fonctions des variables logiques; l'opération de négation est fonction d'une variable logique, et les opérations d'addition logique et de multiplication logique sont fonctions de deux variables logiques et plus.

La particularité de ces fonctions, comme on l'a déjà signalé, consiste dans le fait qu'elles ne peuvent prendre que deux valeurs 0 et 1.

Les fonctions n'admettant que deux valeurs différentes tant pour l'argument que pour la fonction elle-même portent le nom de *fonctions booléennes*. Il est évident qu'il y a aussi d'autres fonctions booléennes, en plus de celles mentionnées. Cependant on peut montrer que le nombre des différentes fonctions booléennes d'un nombre fini de variables logiques est fini.

En effet, supposons qu'on désire former toutes les fonctions booléennes possibles de m variables logiques x, y, z, . . . La totalité des valeurs numériques de ces variables s'appelle *collection*. On peut obtenir au total 2^m collections.

Soit $f(x, y, z, \ldots)$ une fonction booléenne. A chacune des collections correspond une valeur numérique déterminée (1 ou 0) de la fonction booléenne, de sorte qu'on peut considérer la fonction booléenne comme une collection de 2^m variables. Les différentes fonctions booléennes différeront par les valeurs des variables logiques de la collection. Chaque variable n'admettant que deux valeurs, le nombre des différentes fonctions booléennes de m arguments est

$$B(m) = 2^{2^m}. \tag{4-11}$$

Quand $m = 1$, on peut obtenir $2^2 = 4$ diverses fonctions booléennes, à savoir: les constantes 1 et 0, la fonction x elle-même et son inversion $\bar{x}$.

Pour $m = 2$, on compte $2^4 = 16$ diverses fonctions booléennes. Le tableau 4-1 réunit les valeurs de ces fonctions et leurs désignations conventionnelles disposées de telle façon que la seconde moitié du tableau s'obtient de la première par l'opération de négation. En plus des opérations étudiées plus haut, le tableau 4-1 comporte les fonctions suivantes:

$x \oplus y$, addition modulo deux;
$x \to y$, $y \to x$, implication, ou conséquence logique;
$\overline{xy} = x/y$, fonction de Sheffer
$\overline{x + y}$, fonction de Pierce (fonction de Webb);
$\overline{x \oplus y} = x \sim y$, équivalence logique;
$\overline{x \to y}$, $\overline{y \to x}$, fonction d'inhibition.

Il importe de souligner que toutes les fonctions booléennes considérées de deux variables logiques peuvent être établies à l'aide de trois opérations logiques étudiées plus haut, à savoir: négation, multiplication logique et addition logique. Ainsi, on s'assure sans peine, par vérification directe dans toutes les collections (x, y), que les relations suivantes ont lieu:

$$\left.\begin{aligned}
&(a)\quad x \oplus y = \bar{x}y + \bar{y}x;\\
&(b)\quad x \to y = \bar{x} + y;\\
&(c)\quad y \to x = x + \bar{y};\\
&(d)\quad \overline{x \oplus y} = x \sim y = xy + \bar{x}\bar{y};\\
&(e)\quad \overline{x \to y} = x\bar{y};\\
&(f)\quad \overline{y \to x} = y\bar{x}.
\end{aligned}\right\} \tag{4-12}$$

On dit d'une collection d'opérations logiques qu'elle est *complète* si elle permet de représenter n'importe quelle fonction booléenne. En plus de la collection examinée ci-dessus, composée d'opérations de négation, de multiplication logique et d'addition logique, il peut y avoir aussi d'autres collections complètes. Entre autres, la collection complète peut être constituée par une seule opération : fonction de Sheffer $\overline{xy}$, désignée souvent x/y ; on s'en rend compte bien facilement en représentant sous la forme de cette opération les trois opérations mentionnées plus haut de la collection complète

$$\bar{x} = \overline{xx} = x/x\,;$$
$$xy = \overline{\overline{xy}} = \overline{x/y} = (x/y)/(x/y)\,;$$
$$x+y = \overline{\bar{x}\bar{y}} = \bar{x}/\bar{y} = (x/x)/(y/y).$$

Il est facile de voir que la fonction de Pierce représente, elle aussi, une collection complète.

e) Lois et identités de l'algèbre des propositions

Les expressions construites à partir d'un nombre fini de variables logiques $x, y, z, \ldots$, de signes d'opérations logiques de négation, de multiplication logique et d'addition logique, ainsi que des constantes 1 et 0, portent le nom de *formules booléennes* et s'écrivent $A\,(x, y\ z, \ldots)$, $B\,(x, y, z, \ldots)$, ... Pour que les formules booléennes soient comprises sans ambiguïté, il faut mettre chaque nouvelle opération logique entre parenthèses. Or, afin d'éviter des formules trop encombrantes et de minimiser l'emploi des parenthèses, on introduit, par analogie avec l'algèbre élémentaire, un ordre de dominance des opérations : on effectue en premier lieu les opérations de négation, puis celles de multiplication logique et enfin, en dernier lieu, celles d'addition logique.

Chaque formule booléenne peut être considérée comme la représentation d'une certaine fonction booléenne des variables $x, y, z, \ldots$, dont la valeur sur la collection concrète de variables s'obtient sans peine en substituant les valeurs des variables qu'elles prennent sur la collection donnée (0 ou 1) dans la formule booléenne et en effectuant les opérations logiques prescrites.

Pour les mêmes variables logiques, il est possible d'obtenir des formules logiques différentes, parmi lesquelles on rencontre parfois, par exemple, des formules comme A et B qui fournissent l s mêmes valeurs des fonctions booléennes sur toutes les collections identiques de variables logiques $x, y, z, \ldots$, c.-à-d. satisfont à la condition

$$A\,(x, y, z, \ldots) = B\,(x, y, z, \ldots). \qquad (4\text{-}13)$$

Un des problèmes de l'algèbre de Boole consiste précisément à établir des identités de la forme (4-13).

Pour vérifier les identités de l'algèbre de Boole, il suffit de calculer les valeurs des fonctions dans le premier et le second membre de l'identité sur toutes les 2^m collections de variables.

Une autre méthode de vérification de l'identité de deux formules booléennes consiste à définir leurs ensembles de vérité. Si les ensembles de vérité pour deux formules booléennes coïncident, ces formules donnent une seule et même fonction booléenne.

On a démontré au chapitre 1 quelques identités de l'algèbre des ensembles. Chacune de ces identités définit une certaine identité pour les formules booléennes. Considérons par exemple l'identité suivante de l'algèbre des ensembles:

$$(X \cap Y) \cup Z = (X \cup Z) \cap (Y \cup Z).$$

Le premier membre de cette identité est l'ensemble de vérité pour la fonction booléenne $xy + z$, et son second membre l'est pour $(x + z)(y + z)$. Puisque les ensembles de vérité coïncident pour les deux formules booléennes, ces dernières définissent une seule et même fonction booléenne, de sorte que

$$xy + z = (x + z)(y + z).$$

Cette identité est remarquable par le fait qu'elle n'a pas son analogue en algèbre classique, ce qui a été signalé au chapitre 1.

Le lecteur est invité à s'assurer lui-même de l'exactitude des lois et identités suivantes de la logique mathématique en les comparant aux identités correspondantes de l'algèbre des ensembles ou bien en vérifiant les valeurs de vérité dans les premiers et les seconds membres sur toutes les collections de variables logiques.

Lois de l'algèbre de Boole

a) $x + y = y + x$;
b) $xy = yx$; } loi commutative;

c) $(x + y) + z = x + (y + z)$;
d) $(xy)z = x(yz)$; } loi associative;

e) $(x + y)z = xz + yz$;
f) $xy + z = (x + z)(y + z)$. } loi distributive.

Identités de l'algèbre de Boole

a) $x\bar{x} = 0$;
b) $x + \bar{x} = 1$;
c) $x \cdot 1 = x$;
d) $x + 1 = 1$;
e) $x \cdot 0 = 0$;
f) $x + 0 = x$;
g) $xx = x$;
h) $x + x = x$;
i) $\bar{\bar{x}} = x$;
j) $\overline{x + y} = \bar{x}\bar{y}$;
k) $\overline{xy} = \bar{x} + \bar{y}$.

4-3. SYNTHÈSE DES RÉSEAUX COMBINATIONNELS

a) Notion de réseau combinationnel

On entend par *réseau combinationnel* un dispositif technique servant à transformer l'information discrète et ayant n entrées et m sorties. Les signaux fournis aux entrées et prélevés sur les sorties ne peuvent prendre que les valeurs 1 et 0. Ces signaux peuvent être représentés par exemple par un niveau haut et un niveau bas de tension. De cette façon, le réseau combinationnel transforme un mot d'entrée à n lettres de l'alphabet binaire en un mot de sortie à m lettres du même alphabet. On admet que le mot de sortie est prélevé à la sortie au même instant que le mot d'entrée arrive à l'entrée, c.-à-d. que le réseau combinationnel ne comporte pas de retard de signaux.

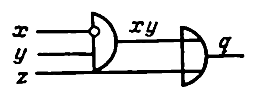


Fig. 4-7. Exemple d'un réseau combinationnel logique

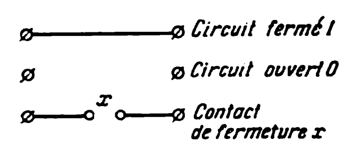
Le réseau combinationnel se construit souvent à partir de circuits logiques élémentaires de négation, de multiplication logique et d'addition logique en reliant en série les sorties d'un circuit logique élémentaire aux entrées d'un autre circuit logique, sans jamais brancher plusieurs sorties sur une seule entrée. Dans le cas où le réseau combinationnel possède un ensemble fini de signaux d'entrée x, y, z, . . . et une seule sortie q, le processus décrit de la mise en série des circuits logiques élémentaires correspond au processus d'établissement de la formule logique $f(x, y, z, \ldots)$ à l'aide des opérations réalisées par les circuits logiques élémentaires utilisés. Ainsi, à la formule logique de la forme

$$q = \bar{x}y + z \tag{4-14}$$

correspondra le réseau combinationnel représenté sur la figure 4-7 et dans le tableau 4-2. Dans bien des cas, il est commode d'assembler le réseau combinationnel avec des éléments réalisant la fonction de Sheffer ou celle de Pierce.

Un des modes de réalisation physique des réseaux combinationnels est la construction de montages à relais et à contacts permettant d'effectuer les opérations logiques par fermeture et ouverture des contacts dans des circuits différents. Désignons par 1 un circuit fermé, et par 0, un circuit ouvert. Pour pouvoir commander l'état du circuit, introduisons des contacts de façon que leur état soit dicté par la valeur des variables d'entrée x, y, z, . . . Il est facile de s'assurer qu'aux opérations de multiplication logique et d'addition logique correspondront alors les couplages des contacts en série et en parallèle, comme il est montré sur la figure 4-8.

En procédant par différents couplages des contacts, on arrive à former des montages réalisant des opérations logiques très variées. On voit sur la figure 4-9 le montage à relais et à contacts réalisant la formule (4-14), et sur la figure 4-10, les montages à relais et à contacts illustrant certaines identités de la logique



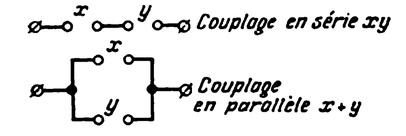


Fig. 4-8. Représentation des opérations logiques élémentaires au moyen des contacts

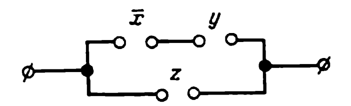


Fig. 4-9. Circuit de commutation pour l'opération $\bar{x}y + z$

mathématique. Au cours de ces dernières années, on réalise de plus en plus souvent les réseaux combinationnels sans contacts

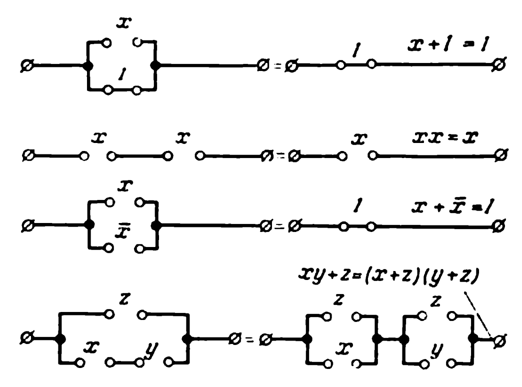


Fig. 4-10. Transformations identiques des circuits de commutation

au moyen d'éléments magnétiques et semi-conducteurs réunis en circuits intégrés.

b) Etablissement de la formule logique correspondant à une table donnée

Lors de la mise au point des réseaux combinationnels, on arrive bien rarement à exprimer les problèmes destinés à être résolus par ce réseau directement sous la forme d'une formule logique; généralement, à la première étape, en utilisant la description verbale des

problèmes, on dresse un tableau établissant la relation entre les valeurs numériques des variables logiques d'entrée et de sortie. Le passage du tableau à la formule logique constitue la seconde étape de la synthèse des réseaux combinationnels.

Un tableau est dit *élémentaire* s'il correspond au cas de la multiplication logique ou de l'addition logique des propositions initiales. Si le tableau correspond à l'opération de multiplication logique, on verra des zéros dans toutes les lignes de la colonne q, sauf dans une seule ligne, qui correspond aux propositions initiales vraies: dans cette ligne on verra 1. Si le tableau correspond à l'opération d'addition logique, on verra des unités dans toutes les lignes de la colonne q sauf dans la seule ligne, correspondant aux propositions initiales fausses, où il y aura zéro.

Explicitons le mode d'obtention de la formule logique pour un tableau élémentaire sur les exemples des tableaux 4-3 et 4-4.

Dans le tableau 4-3, le zéro ne figure que dans la deuxième ligne de la colonne q, ce qui correspond à l'opération d'addition logique. Puisque dans cette ligne $\bar{x} = 0$ et $y = 0$, on a $q = \bar{x} + y$.

Tableau 4-3

$q = \bar{x} + y$

xy	q
0 0	1
1 0	0
0 1	1
1 1	1

Tableau 4-4

$q = x\bar{y}$

xy	q
0 0	0
1 0	1
0 1	0
1 1	0

Dans le tableau 4-4 l'unité ne figure que dans la deuxième ligne de la colonne q, ce qui correspond à l'opération de multiplication logique. Puisque dans cette ligne $x = 1$ et $\bar{y} = 1$, on a $q = x\bar{y}$.

Si le tableau n'est pas élémentaire, on établit la formule logique correspondante par un des deux procédés standard suivants.

Premier procédé. Dans la colonne q, il faut remplacer successivement toutes les unités sauf une par des zéros, puis établir les formules logiques pour chacun des tableaux élémentaires ainsi formés et prendre leur somme logique.

Deuxième procédé. Dans la colonne q, il faut remplacer successivement tous les zéros sauf un par des unités, puis établir les formules logiques pour chacun des tableaux élémentaires ainsi formés et prendre leur produit logique.

Illustrons ces procédés sur l'exemple d'obtention de la formule logique pour le tableau 4-5. D'après le premier procédé, considérons

indépendamment les tableaux où l'unité ne figure que dans la deuxième et la troisième ligne de la colonne q. Les formules logiques correspondantes sont respectivement $\bar{x}y$ et $x\bar{y}$. Prenant leur somme logique, on a :

$$q = \bar{x}y + x\bar{y}.$$

D'après le deuxième procédé, considérons séparément les tableaux contenant les zéros seulement dans la première et la dernière ligne de la colonne q. Les formules logiques qui correspondent à ces tableaux sont respectivement $x + y$ et $\bar{x} + \bar{y}$. Prenons le produit logique de ces deux formules ; il vient :

$$q = (x + y)(\bar{x} + \bar{y}).$$

Tableau 4-5

Addition modulo deux

xy	q
0 0	0
0 1	1
1 0	1
1 1	0

On voit que le premier et le deuxième procédé fournissent des formules booléennes différentes pour un seul et même tableau. Or, utilisant les identités de la logique mathématique, on arrive sans peine à ramener la deuxième formule à la première. Ouvrant les parenthèses et se rappelant que $x\bar{x} = 0$ et $y\bar{y} = 0$, on obtient :

$$q = x\bar{x} + x\bar{y} + y\bar{x} + y\bar{y} = x\bar{y} + y\bar{x}.$$

La méthode proposée est universelle, car elle permet d'obtenir la formule logique et de construire ensuite le réseau combinationnel pour n'importe quel tableau. Cependant cette méthode fournit le plus souvent des formules très encombrantes, complexes, qui ne sont réalisables que moyennant une grande quantité de divers éléments. La *complexité* de la formule traduit le nombre d'opérations qu'elle recèle, de sorte que la complexité de la formule $\bar{x}$ est le nombre 1, et celle de la formule $(\bar{x} + y)(\bar{y} + z)$, le nombre 5 (deux négations, deux additions, une multiplication). En qualité d'exemple, établissons la formule logique d'après le premier procédé pour le tableau 4-2 :

$$q = \bar{x}y\bar{z} + \bar{x}\bar{y}z + x\bar{y}z + \bar{x}yz + xyz. \qquad (4\text{-}15)$$

Cette formule est beaucoup plus complexe que (4-14) bien qu'elles correspondent au même tableau. On voit donc que, pour la synthèse des réseaux combinationnels, il est très important de simplifier ou de minimiser les formules booléennes.

c) Simplification des formules booléennes

Pour simplifier les formules booléennes, on utilise les identités de la logique mathématique étudiées dans le paragraphe précédent. Il est vrai que leur emploi efficace exige de l'habitude et de l'art dans leur manipulation, qui ne s'acquièrent qu'après une certaine expérience de pareilles transformations. Or, il existe quelques procédés standard qui permettent de mener à bien, dans la plupart des cas, la simplification d'une formule compliquée [24, 30].

La simplification d'une formule booléenne commence par la recherche de l'une des formes suivantes: $A\bar{B} + AB$, $A + AB$, $A + \bar{A}B$, où A et B symbolisent soit les variables logiques ellesmêmes, soit les produits logiques de plusieurs variables. Chacune des expressions obtenues peut être écrite sous une forme plus simple:

$$A\bar{B} + AB = A(\bar{B} + B) = A; \qquad (4\text{-}16)$$

$$A + AB = A(1 + B) = A; \qquad (4\text{-}17)$$

$$A + \bar{A}B = (A + AB) + \bar{A}B = A + B. \qquad (4\text{-}18)$$

Reprenons la formule (4-15) et essayons de la simplifier par application des identités étudiées. Groupant le premier et le quatrième terme, puis le troisième et le cinquième et appliquant l'identité (4-16), on obtient:

$$q = \bar{x}y + \bar{x}\bar{y}z + xz.$$

La simplification se poursuit sans difficulté:

$$q = \bar{x}(y + \bar{y}z) + xz = \bar{x}(y + z) + xz = \bar{x}y + \bar{x}z + xz = \bar{x}y + z.$$

En simplifiant les formules logiques, il ne faut pas oublier l'identité $A + A = A$, d'où il s'ensuit que chacun des termes peut être utilisé plusieurs fois dans les combinaisons avec d'autres termes. Compte tenu de ce fait, simplifions la formule suivante:

$$q = x\bar{y}\bar{z} + xy\bar{z} + \bar{x}yz + xyz.$$

Appliquant l'identité (4-16) aux deuxième et quatrième termes, ainsi qu'aux troisième et quatrième, on obtient:

$$q = x\bar{y}\bar{z} + xy + yz = x(y + \bar{y}\bar{z}) + yz = x(y + \bar{z}) + yz = xy + x\bar{z} + yz.$$

Les identités susmentionnées ne permettent pas de pousser la simplification plus loin. La question se pose de savoir si la forme donnée est élémentaire. Pour le vérifier, voyons si elle ne comporte pas de termes superflus. On dit que le terme est superflu si dans chaque collection de variables où ce terme devient l'unité il y a encore un groupe d'autres termes qui devient l'unité de même. C'est ainsi que le terme xy devient l'unité dans la collection $x = 1$, $y = 1$. La somme

de deux autres termes donne $x\bar{z} + yz = \bar{z} + z = 1$. Ainsi donc, le terme xy est superflu et peut être éliminé de la formule, ce qui donne en définitive

$$q = x\bar{z} + yz.$$

De toutes les formes ne contenant aucun terme superflu, en les examinant toutes une à une, on recherche la forme élémentaire. Il peut y avoir plusieurs formes de cette espèce.

Pour les cas où le nombre de variables d'entrée n'est pas supérieur à quatre, il est commode de rechercher les formes élémentaires des formules booléennes en employant des tableaux spéciaux dits *cartes de Karnaugh* ou *diagrammes de Veitch*. Le tableau 4-6 repré-

Tableau 4-6

Carte de Karnaugh pour la fonction booléenne de trois variables

xy	z = 0	z = 1
0 0	1	2
0 1	3	4
1 1	5	6
1 0	7	8

Tableau 4-7

Carte de Karnaugh pour la fonction booléenne définie par le tableau 4-2

xy	z = 0	z = 1
0 0		×
0 1	×	×
1 1		×
1 0		×

sente la carte de Karnaugh pour la fonction booléenne de trois variables. Chacune des cases de cette carte (numérotées pour faciliter l'exposé) correspond à une collection de variables x, y, z. Les cases correspondant aux collections dans lesquelles la fonction booléenne admet la valeur 1 sont marquées par des croix, comme il est montré sur le tableau 4-7 représentant la carte de Karnaugh pour la fonction booléenne définie par le tableau 4-2. La carte est construite de telle façon que les zones correspondant aux divers groupes de cases voisines peuvent être considérées comme les ensembles de vérité pour les propositions correspondant aux formules booléennes élémentaires à une et à deux variables (les bords supérieur et inférieur de la carte sont considérés comme étant collés ensemble, de sorte que les cases les plus hautes et les plus basses sont voisines). De cette façon, aux zones formées par quatre cases voisines correspondent les formules booléennes d'une variable: 1234 — $\bar{x}$; 5678 — x; 3456 — y; 1278 — $\bar{y}$; 1357 — $\bar{z}$; 2468 — z.

Aux zones de deux cases voisines correspondent les fonctions booléennes de deux variables de la forme: 12 — $\bar{x}\bar{y}$; 56 — xy; 68 — xz, et ainsi de suite.

Pour rechercher la formule booléenne déterminée par une configuration quelconque de cases de la carte de Karnaugh, il suffit de représenter cette combinaison sous la forme de réunion des zones indiquées plus haut. La plus élémentaire des représentations de cette espèce traduit justement la formule booléenne la plus simple. Par exemple, dans le cas du tableau 4-7, la zone marquée peut être représentée comme la réunion de la bande verticale 2468, ou z, avec la bande horizontale 34, ou $\bar{x}y$, d'où il vient $q = z + \bar{x}y$.

Lorsqu'il s'agit de simplifier les formules booléennes à un grand nombre de variables, l'emploi des méthodes proposées devient pénible. Dans ces cas il est rationnel d'employer des méthodes spéciales de minimisation, parmi lesquelles on conseille tout spécialement la méthode de Quine-McCluskey [31].

d) Exemples de synthèse des réseaux combinationnels

Exemple 4-2. Construire le circuit logique de l'additionneur à une position à deux entrées (*semi-additionneur*).

L'additionneur à une position effectue l'addition de deux nombres à un chiffre d'après la règle

$$0 + 0 = 0,\ 0 + 1 = 1 + 0 = 1,\ 1 + 1 = 1\ 0.$$

Dans le dernier cas on laisse 0 à la position donnée, et l'unité est reportée à la position suivante. Désignons par x et y les valeurs de la position donnée de termes, par q la valeur de la position donnée de la somme et par p l'unité reportée à la position suivante. Le fonctionnement du semi-additionneur sera résumé alors par le tableau 4-8, d'où l'on tire:

$$q = x\bar{y} + \bar{x}y\,;\quad p = xy. \tag{4-19}$$

Le circuit logique du semi-additionneur est donné sur la figure 4-11.

Tableau 4-8

Semi-additionneur

xy	q	p
0 0	0	0
0 1	1	0
1 0	1	0
1 1	0	1

Tableau 4-9

Décodeur

xy	q_1	q_2	q_3	q_4
0 0	1	0	0	0
0 1	0	1	0	0
1 0	0	0	1	0
1 1	0	0	0	1

Exemple 4-3. Circuit de sélection (*décodeur*). Un circuit logique à deux entrées x, y et à quatre sorties q_1, q_2, q_3, q_4 est appelé circuit de sélection si le signal 1 n'apparaît qu'à une seule sortie pour chacune des quatre combinaisons possibles

de signaux d'entrée. A ces conditions satisfait le tableau 4-9 dans lequel les signaux de sortie du circuit sont

$$q_1=\bar{x}\bar{y}; \quad q_2=\bar{x}y; \quad q_3=x\bar{y}; \quad q_4=xy. \tag{4-20}$$

On voit sur la figure 4-12 la structure logique du circuit de sélection.

Exemple 4-4. Dispatcher automatique. Donnons un cas très simplifié de construction d'un dispatcher automatique commandant l'atterrissage des avions sur l'aérodrome. Considérons l'aérodrome à piste d'atterrissage unique. L'avion qui vient donne ses indicatifs d'appel, et le dispatcher automatique délivre l'autorisation d'atterrir. S'il y a plusieurs avions qui viennent en même temps, le dispatcher automatique donne à un avion l'autorisation d'atterrir et propose aux autres de rester un certain temps dans l'air.

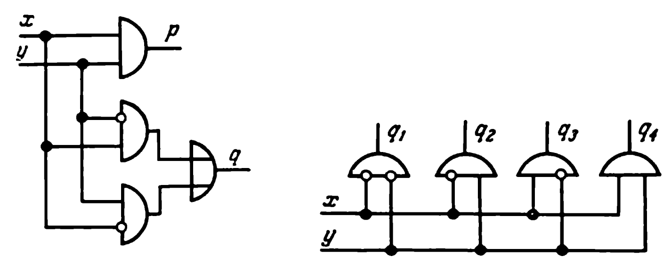


Fig. 4-11. Semi-additionneur   Fig. 4-12. Décodeur

Attribuons à tous les avions de l'aérodrome des numéros de 1 à n. Si plusieurs avions se présentent en même temps, la préférence est accordée à l'avion de numéro plus petit. On fera abstraction du cas où il y a plus de trois avions sollicitant l'autorisation d'atterrissage; on admet que ce cas est impossible. Il s'agit de construire un réseau combinationnel remplissant les fonctions de dispatcher automatique.

Donnons la formulation du problème en termes de logique mathématique. Désignons par x_k le signal d'arrivée du k-ième avion. Le réseau combinationnel logique doit délivrer au k-ième avion un des trois signaux suivants:

q_k^0, atterrir immédiatement;

q_k', faire un tour au-dessus de l'aérodrome;

q_k'', faire deux tours au-dessus de l'aérodrome.

Il est difficile de décrire le fonctionnement du réseau par un tableau; procédons donc autrement. Remarquons que les signaux q_k^0, q_k' et q_k'' ne peuvent prendre la valeur unité que si c'est le k-ième avion qui se présente, donc si $x_k = 1$. Par conséquent, le signal x_k doit figurer à titre de facteur dans les formules logiques de q_k^0, q_k' et q_k'', ces dernières ayant donc la forme

$$q_k^0=A_k^0x_k, \quad q_k'=A_k'x_k, \quad q_k''=A_k''x_k. \tag{4-21}$$

On obtient les formules logiques de A_k^0, A_k' et A_k'' en effectuant le passage de k à $k+1$.

La condition $A_{k+1}^0 = 1$ est vérifiée en l'absence de tout signal de l'ensemble $x_1, \ldots, x_k$, c.-à-d. en l'absence de tout signal de l'ensemble $x_1, \ldots, x_{k-1}$ et en l'absence de signal x_k. Ainsi donc,

$$A_{k+1}^0=A_k^0\bar{x}_k. \tag{4-22}$$

La condition $A_{k+1}' = 1$ est vérifiée en présence d'un signal de l'ensemble $x_1, \ldots, x_k$, c.-à-d. en présence d'un signal de l'ensemble $x_1, \ldots, x_{k-1}$ et en l'ab-

sence de signal x_k ou en l'absence de tout signal de l'ensemble $x_1, \ldots, x_{k-1}$ et en présence de signal x_k. Ainsi donc,

$$A'_{k+1} = A'_k \bar{x}_k + A^0_k x_k = A'_k \bar{x}_k + q^0_k. \tag{4-23}$$

La condition $A''_{k+1} = 1$ est vérifiée s'il y a deux signaux de l'ensemble $x_1, \ldots, x_k$, c.-à-d. en présence de deux signaux de l'ensemble $x_1, \ldots, x_{k-1}$ et en

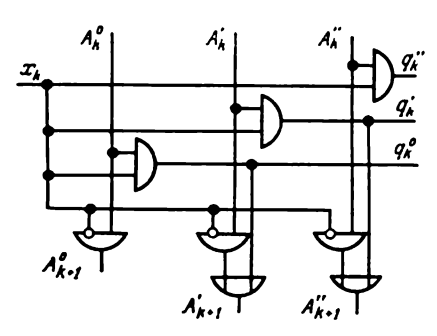


Fig. 4-13. Dispatcher automatique

l'absence de x_k ou en présence d'un signal de l'ensemble $x_1, \ldots, x_{k-1}$ et en présence de x_k. Ainsi donc,

$$A''_{k+1} = A''_k \bar{x}_k + A'_k x_k = A''_k \bar{x}_k + q'_k. \tag{4-24}$$

Les formules (4-21) à (4-24) décrivent complètement la structure logique d'une cellule du réseau combinationnel commandant l'atterrissage du k-ième avion; on voit son schéma sur la figure 4-13. Construisant de pareils réseaux pour chaque cellule et les groupant, on obtient le schéma complet du dispatcher automatique.

4-4. NOTION D'AUTOMATES FINIS

a) Réseau combinationnel comme automate fini sans mémoire

On fait usage du terme « automate fini » pour désigner une classe de systèmes dynamiques numériques employés en automatique, télémécanique et technique de calcul [25, 31].

Nous rencontrons les automates finis, sous la forme la plus élémentaire, en étudiant les réseaux combinationnels tels qu'on vient de les considérer; il est commode toutefois de leur donner une définition quelque peu plus générale.

Soient plusieurs variables logiques $u_1, \ldots, u_r$ telles que leur totalité puisse être considérée comme une grandeur à plusieurs dimensions $u = (u_1, \ldots, u_r)$. Soit ensuite f la fonction logique de ces variables constituant une proposition logique composée

$$q = f(u_1, \ldots, u_r). \tag{4-25}$$

La variable logique q peut être considérée comme la variable de sortie d'un réseau combinationnel réalisant l'opération logique f sur les variables $u_1, \ldots, u_r$.

En pratique, les réseaux combinationnels possèdent généralement, au lieu d'une sortie unique q, plusieurs sorties $q_1, \ldots, q_m$, de sorte que chaque sortie réalise sa propre fonction logique des mêmes variables logiques

$$\left.\begin{array}{l} q_1 = f_1(u_1, \ldots, u_r); \\ \ldots\ldots\ldots\ldots \\ q_m = f_m(u_1, \ldots, u_r). \end{array}\right\} \qquad (4\text{-}26)$$

La structure de ce réseau combinationnel est montrée sur la figure 4-14.

Les relations (4-26) peuvent avoir une forme plus condensée si l'on fait intervenir une variable logique de sortie à plusieurs dimensions $q = (q_1, \ldots, q_m)$ et une fonction logique à plusieurs dimensions $f = (f_1, \ldots, f_m)$. L'ensemble des fonctions (4-26) s'écrit alors sous la forme d'une fonction multidimensionnelle unique

$$q = f(u). \qquad (4\text{-}27)$$

Supposons ensuite que le réseau combinationnel réalise instantanément son opération logique, c.-à-d. que, pour une certaine valeur $u \in U$ de la variable d'entrée arrivée dans le réseau en un certain instant, on obtient à la sortie la valeur correspondante $q = f(u)$ pratiquement au même instant.

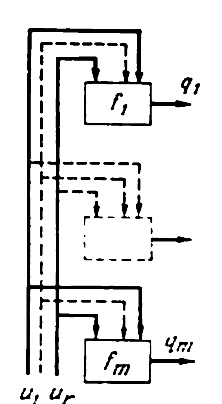


Fig. 4-14. Automate fini sans mémoire

Fig. 4-15. Echelle du temps discret

Pour donner la description du fonctionnement du réseau combinationnel dans le temps, on introduit la notion de *temps discret*. Sur l'axe de temps, marquons les instants, où la variable d'entrée peut subir des modifications, en les désignant par $t_0, t_1, t_2, \ldots$ Il est commode de désigner les instants notés simplement par des entiers non négatifs 0, 1, 2, . . .; on les appelle *unités de temps*. On voit se former alors une échelle (fig. 4-15) qui porte le nom d'échelle du temps discret.

Désignons par la lettre n l'unité de temps actuelle, c.-à-d. correspondant à l'instant de temps actuel. Les valeurs des variables d'entrée et de sortie à l'instant actuel seront désignées par $u\ [n]$ et $q\ [n]$. Les valeurs prises par les variables d'entrée et de sortie k unités avant l'instant actuel seront désignées par $u\ [n-k]$ et $q\ [n-k]$. Mettant en équation le fonctionnement du réseau combinationnel dans le temps, on écrit

$$q\ [n] = f\ (u\ [n]). \qquad (4\text{-}28)$$

Les réseaux combinationnels constituent la forme la plus élémentaire d'automates finis, appelés *automates finis sans mémoire.* Par cette appellation on souligne le fait que la variable de sortie du réseau combinationnel à chaque unité de temps dépend uniquement des signaux d'entrée à cette même unité de temps et ne dépend ni de l'état du réseau, ni des signaux d'entrée pendant les unités de temps précédentes.

b) Automates finis de la forme générale

Etant la forme la plus élémentaire des automates finis, les automates finis sans mémoire ne peuvent servir comme éléments constitutifs de dispositifs de calcul et de systèmes automatiques que d'une façon assez limitée. On élargit avantageusement les possibilités des automates finis en leur ajoutant des éléments complémentaires effectuant le retardement des signaux d'entrée d'une unité de temps : ce sont les *éléments de mémoire.* On en voit une représentation schématique sur la figure 4-16. Combinant les éléments de mémoire avec les réseaux combinationnels, on obtient les automates finis de la forme générale.

Fig. 4-16. Cellule de mémoire

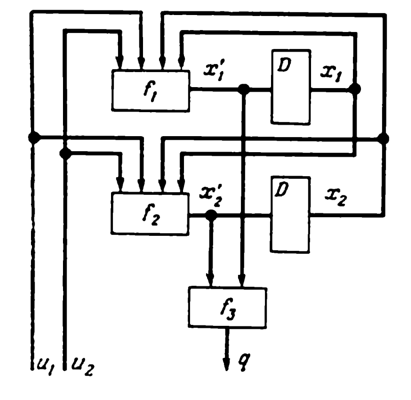


Fig. 4-17. Automate fini avec mémoire

Un automate fini pourvu de mémoire se caractérise par le fait que son état ne se définit pas seulement par la nature des signaux d'entrée à l'unité de temps donnée mais aussi par l'état dans lequel il s'est trouvé à l'unité de temps précédente. On fait généralement une distinction entre l'état de l'automate et son signal à la sortie.

Le schéma représenté sur la figure 4-17 peut servir d'exemple d'un automate fini avec mémoire. Il comprend trois réseaux combinationnels réalisant trois fonctions logiques f_1, f_2 et f_3 et deux éléments de mémoire. Sa particularité consiste en ce que les signaux x_1 et x_2, recueillis à la sortie de deux réseaux combinationnels, arrivent, avec un retard d'une unité de temps, à l'entrée des mêmes réseaux. Ces signaux circulant dans le circuit de l'automate fini peuvent être appelés variables décrivant l'état de l'automate. Le signal de sortie q est une fonction des variables d'état.

Donnons la description du fonctionnement du montage représenté à l'unité de temps n. Pour simplifier l'écriture, omettons d'indiquer le numéro n de l'unité de temps. Les valeurs prises par les variables d'état à l'unité de temps donnée à l'entrée de la cellule de mémoire seront marquées par un apostrophe. Les signaux x'_1, x'_2 et q ont pour équations

$$\left.\begin{aligned} x'_1 &= f_1(x_1, x_2, u_1, u_2); \\ x'_2 &= f_2(x_1, x_2, u_1, u_2); \\ q &= f_3(x'_1, x'_2) = \varphi(x_1, x_2, u_1, u_2). \end{aligned}\right\} \quad (4\text{-}29)$$

Pour abréger l'écriture, faisons intervenir des grandeurs multidimensionnelles $u = (u_1, u_2)$, $x = (x_1, x_2)$ et une fonction multidimensionnelle $f = (f_1, f_2)$. Les équations (4-29) prennent alors la forme

$$x' = f(x, u); \quad q = \varphi(x, u). \quad (4\text{-}30)$$

La fonction $f(x, u)$ définissant le nouvel état intérieur de l'automate fini porte le nom de *fonction des transitions*, et la fonction $\varphi(x, u)$ définissant la nouvelle valeur du signal de sortie s'appelle *fonction des sorties*.

Désignant par U l'ensemble des signaux d'entrée, par Q l'ensemble des signaux de sortie, par X l'ensemble des états, on arrive à donner à l'automate fini une définition formelle suivante.

On entend par *automate fini* la collection des cinq grandeurs suivantes:

$$A = (U, Q, X, f, \varphi), \quad (4\text{-}31)$$

où $f: X \times U \to X$ est la fonction des transitions et $\varphi: X \times U \to Q$ la fonction des sorties.

D'habitude, on définit les automates finis non par des équations (4-30) mais par deux tableaux, appelés respectivement *tableau des transitions* et *tableau des sorties*. Les lignes des deux tableaux correspondent aux différents signaux à l'entrée de l'automate fini, et les colonnes, à ses différents états. A l'intersection des lignes et des colonnes. on trouve dans le tableau des transitions les valeurs de la fonction $f(x, u)$, et dans le tableau des sorties, celles de la fonction $\varphi(x, u)$. Les tableaux 4-10 et 4-11 donnent l'exemple de définition

Tableau 4-10

Tableau des transitions de l'automate fini

u	x		
	1	2	3
a	2	3	3
b	3	2	2

Tableau 4-11

Tableau des sorties de l'automate fini

u	x		
	1	2	3
a	c	c	d
b	d	c	c

d'un automate fini admettant deux signaux possibles à l'entrée a et b, deux signaux possibles à la sortie c et d et susceptible de trois états différents 1, 2 et 3.

Il y a un autre mode de définition d'automates finis, qui est plus instructif: c'est la définition au moyen de graphes orientés. Les sommets du graphe, représentés au moyen de petits cercles, désignent les différents états de l'automate. Deux sommets x_i et x_k sont reliés par un arc allant de x_i vers x_k dans le cas où il y a un signal u qui fait passer l'automate de l'état x_i à l'état x_k. Chaque arc est marqué par le signal u commandant la transition donnée, et aussi, le plus souvent, par le signal q (entre parenthèses) qu'on obtient à la sortie après la transition donnée. Le graphe correspondant aux tableaux 4-10 et 4-11 est donné à la figure 4-18.

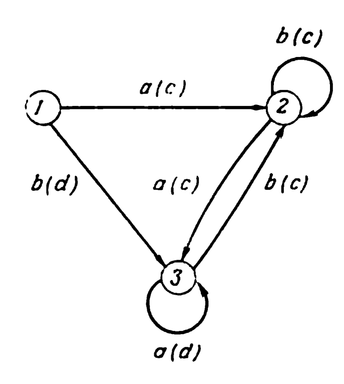


Fig. 4-18. Graphe des transitions d'un automate fini

Les automates finis représentent le modèle mathématique d'une classe très étendue de systèmes dynamiques fonctionnant en temps discret, parmi lesquels se rangent aussi quelques systèmes dynamiques à étapes multiples étudiés dans la deuxième partie de cet ouvrage. La théorie des automates finis ne saurait être exposée ici en détail: quelques exemples, empruntés à [32], nous suffiront.

Exemple 4-5. Générateur d'unités. On voit à la figure 4-19, *a* le graphe d'un automate fini qui, à chaque unité de temps, délivre le signal 1 sans liaison avec le signal à l'entrée, une fois que le signal 1 a été reçu à son entrée. Les fonctions des transitions $x' = f(x, u)$ et des sorties $q = \varphi(x, u)$ de cet automate sont données dans le tableau 4-12, d'où il vient:

$$x' = q = u + x. \tag{4-32}$$

La structure de cet automate est représentée sur la figure 4-19,*b*; on y remarque une entrée auxiliaire r grâce à laquelle on fait cesser la génération d'unités en y faisant parvenir un signal.

Exemple 4-6. Compteur. On voit à la figure 4-20,*a* le graphe d'un automate fini délivrant à la sortie le signal 1 chaque fois qu'un couple d'unités, en alternance avec un nombre arbitraire de zéros, a été appliqué à l'entrée. Un tel dispositif porte le nom de *compteur à base deux*. La fonction des transitions et la fonction des sorties de cet automate sont données par le tableau 4-13 d'où on trouve:

$$x' = u\bar{x} + \bar{u}x\,;\quad q = ux. \tag{4-33}$$

La structure du compteur à base deux est donnée à la figure 4-20,*b*. On y remarque une entrée auxiliaire r: appliquant un signal à cette entrée, on

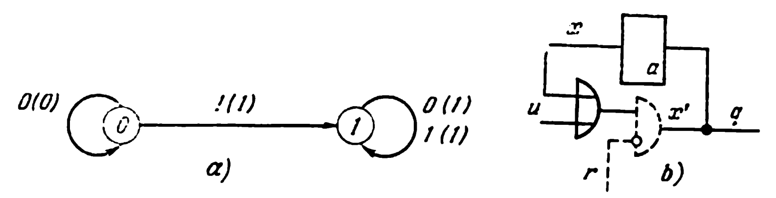


Fig. 4-19. Graphe des transitions et structure d'un générateur d'unités

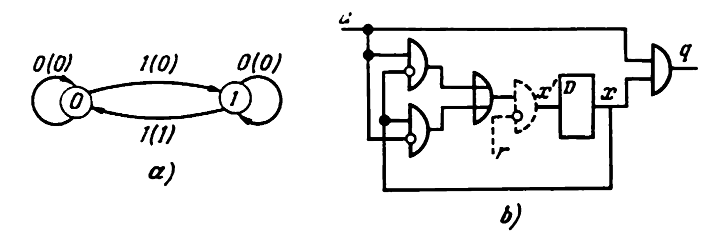


Fig. 4-20. Graphe des transitions et structure d'un compteur binaire

arrête le compteur et on le fait passer à l'état initial 0. La représentation conventionnelle d'un compteur à base deux est donnée à la figure 4-21,*a*. A la figure 4-21,*b* on voit le schéma de n compteurs à base deux couplés en série et formant un compteur à base 2^n.

Tableau 4-12

Générateur d'unités

ux	x'	q
0 0	0	0
1 0	1	1
0 1	1	1
1 1	1	1

Tableau 4-13

Compteur binaire

ux	x'	q
0 0	0	0
1 0	1	0
0 1	1	0
1 1	0	1

Exemple 4-7. Modèle du processus d'apprentissage. A la base d'un compteur à base deux, on construit un schéma susceptible de simuler, d'une manière très simplifiée il est vrai, le processus de l'apprentissage. Ce schéma est représenté à la figure 4-22. Le système en question possède deux entrées a et b et

apprend que l'arrivée du signal unité à l'entrée a est suivie de l'arrivée du signal unité à l'entrée b. Si cela se reproduit 2^n fois (pas obligatoirement de suite et même, peut-être, avec beaucoup de fautes), le schéma apprend à prévoir l'unité à l'entrée b chaque fois que l'unité arrive à l'entrée a, ce qui s'exprime par le fait que le schéma délivre l'unité à la sortie q toutes les fois que l'unité apparaît à l'entrée a.

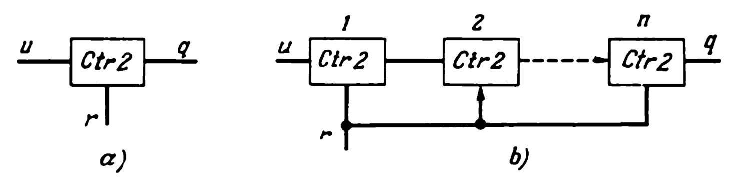


Fig. 4-21. Compteur binaire et compteur 2^n

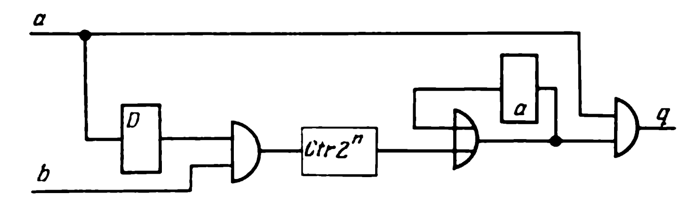


Fig. 4-22. Modèle du processus d'apprentissage

Le schéma de la figure 4-22 simule un processus de l'apprentissage très rudimentaire. Moyennant quelques artifices, on arrive à perfectionner le schéma de telle façon que le fait « a suivi de b » ne soit retenu que si cet événement a eu lieu 2^n fois sans une seule faute, et que ce fait soit ignoré en cas de réalisation

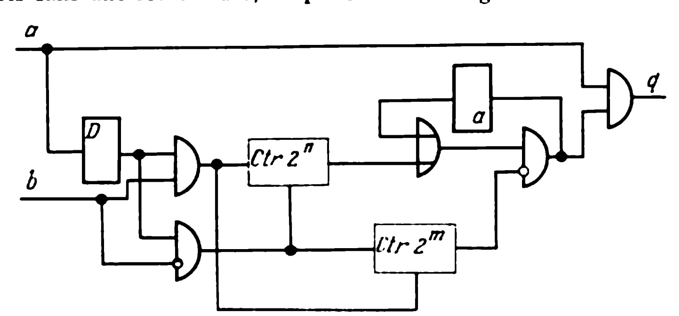


Fig. 4-23. Modèle perfectionné du processus d'apprentissage

de l'événement « a n'est pas suivi de b » 2^m fois sans interférence de l'événement « a suivi de b ». Ce schéma est représenté à la figure 4-23.

PROBLÈMES AU CHAPITRE 4

4-1. Quelle opération logique décrit le fonctionnement du circuit de commande de l'exemple 4-1?

4-2. Quelle est la différence entre la formule booléenne et la fonction booléenne?

4-3. A l'aide des circuits de commutation, vérifier les identités $x + \bar{x} = 1$; $x + 1 = 1$; $x \cdot 0 = 0$; $xx = x$.

4-4. Etablir le réseau d'éclairage d'un local possédant deux portes aux extrémités opposées, permettant de brancher l'éclairage en entrant par une porte quelconque et de le débrancher en sortant par une porte quelconque.

4-5. Simplifier les formules booléennes:

$$xyz + x\bar{y}z + xy\bar{z}v\ ;$$

$$\bar{x}y\bar{z} + x\bar{y}\bar{z} + xy\bar{z}\ ;$$

$$xy + z + \overline{(xy + z)}\,(zv + x).$$

4-6. Construire le circuit logique d'un additionneur binaire à une position à trois entrées en désignant par x, y les valeurs des positions des nombres à additionner; par z, l'unité reportée de la position précédente de la somme; par p, l'unité reportée à la position suivante; par q, la valeur de la position donnée de la somme.

4-7. Etablir le circuit logique d'un additionneur binaire à une position à trois entrées en utilisant des additionneurs binaires à une position à deux entrées.

4-8. Utilisant les résultats des problèmes 4-6 et 4-7, construire le circuit logique d'un additionneur binaire à positions multiples.

CHAPITRE 5

ASPECTS ENSEMBLISTES DE LA THÉORIE DES PROBABILITÉS ET ÉLÉMENTS DE STATISTIQUE MATHÉMATIQUE

5-1. NOTION DE PROBABILITÉ

a) Evénement. Espace des épreuves

A la base de la théorie des probabilités est mis le concept d'*événement aléatoire*. Nous associons ce concept à la réalisation d'une certaine expérience.

En parlant d'une expérience, on s'attache à apprendre son résultat ou, comme on dit en théorie des probabilités, le *cas* ou l'*épreuve* qui se réalise à la suite de cette expérience. Admettons que dans les conditions données l'expérience donne lieu à un nombre fini d'épreuves $z_1, \ldots, z_m$: dont la collection complète sera désignée par

$$Z = \{z_1, \ldots, z_m\}. \tag{5-1}$$

Il est évident que toute expérience doit obligatoirement donner lieu à une certaine épreuve $z \in Z$; d'autre part, aucune expérience ne peut avoir pour résultat deux épreuves ou plus. L'ensemble Z sera appelé *espace des épreuves* liées à l'expérience, et ses éléments $z \in Z$, *événements élémentaires* dans l'espace Z.

Or, ce ne sont pas les événements élémentaires eux-mêmes qui nous préoccupent surtout, mais leurs collections déterminées, qui représentent des sous-ensembles de Z. Tout sous-ensemble S de l'ensemble Z sera appelé *événement* dans l'espace des épreuves Z:

$$S \subseteq Z. \tag{5-2}$$

En disant que l'événement S *se produit* ou *se réalise*, on sous-entend que l'événement élémentaire z, qui est l'épreuve liée à l'expérience, est contenu dans S.

Pour deux événements arbitraires S_1 et S_2 appartenant à l'espace des épreuves Z, les définitions suivantes sont possibles:

La *réunion* $S_1 \cup S_2$ des événements S_1 et S_2: c'est l'événement qui consiste dans la réalisation d'au moins un des événements S_1 et S_2.

L'*intersection* $S_1 \cap S_2$ des événements S_1 et S_2: c'est l'événement qui consiste dans la réalisation de S_1 et de S_2. Les événements S_1 et S_2 sont dits *incompatibles* si la réalisation de l'un d'eux rend impossible la réalisation de l'autre, c.-à-d. si $S_1 \cap S_2 = \varnothing$.

Le *complémentaire* $\overline{S}$ de l'événement S est l'événement consistant dans la non-réalisation de S.

L'*événement certain* consiste dans la réalisation d'au moins un des événements de l'espace Z.

L'*événement impossible* est l'ensemble vide $\varnothing$ qui évoque la non-réalisation d'aucun des événements de Z.

Exemple 5-1. On jette deux dés à jouer à faces numérotées de 1 à 6. Considérons l'ensemble $J = \{1, 2, \ldots 6\}$. L'épreuve qui se réalise à la suite de l'expérience représente un couple ordonné $z = (i, k)$ où $i, k \in J$, i étant le nombre de points du premier dé et k celui du deuxième. Il est commode de représenter les épreuves sous la forme de points du plan (i, k), comme il est montré sur la figure 5-1. On voit que l'espace des épreuves se compose de 36 points $Z = \{(1,1), (1,2), \ldots, (6,6)\}$. Les événements possibles sont très variés ici. Désignons par S_j l'événement de l'apparition de j points. Alors $S_2 = \{(1,1)\}$; $S_7 = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$; $S_{11} = \{(6,5), (5,6)\}$, $S_{12} = \{(6,6)\}$, etc.

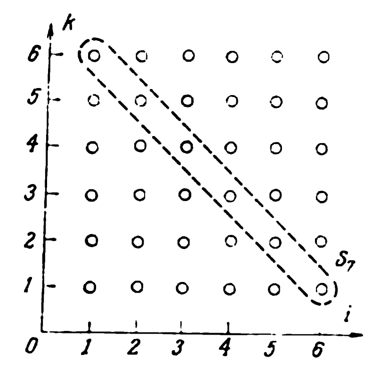


Fig. 5-1. Espace des épreuves pour le cas où l'on jette deux dés

Les événements qui peuvent se produire sur l'espace fini des épreuves Z complété de l'ensemble vide $\varnothing$ forment la classe des événements $\mathfrak{F}$ appelée *classe finie additive*, ou *corps booléen*, ou *algèbre de Boole*. Elle satisfait aux conditions suivantes:

si $S \in \mathfrak{F}$, alors $\overline{S} \in \mathfrak{F}$;

si $S_1 \in \mathfrak{F}$ et $S_2 \in \mathfrak{F}$, alors $S_1 \cup S_2 \in \mathfrak{F}$.

Appliquant ces deux conditions l'une après l'autre, on montre que la réunion ou l'intersection d'un nombre fini arbitraire d'ensembles S_k du corps booléen $\mathfrak{F}$, ainsi que la différence de tels ensembles, appartiennent à $\mathfrak{F}$.

Si l'espace Z contient un nombre fini d'éléments, la classe des événements possibles dans cet espace sera aussi finie. Il est commode de définir alors une classe $\mathfrak{F}_0$ d'ensembles dans l'espace Z, qui sera appelée classe initiale. Appliquant les conditions mentionnées un nombre fini de fois aux ensembles de $\mathfrak{F}_0$, on obtient la classe complète des événements vérifiant ces conditions, c.-à-d. un corps booléen $\mathfrak{F}$. On dit alors que le corps booléen $\mathfrak{F}$ est engendré par la classe initiale $\mathfrak{F}_0$. Il y a de nombreux modes de choix de la classe initiale $\mathfrak{F}_0$ et chacun d'eux donne lieu à un corps booléen particulier.

Exemple 5-2. Si $\mathfrak{F}_0 = Z$, alors $Z \in \mathfrak{F}$. Donc, $\overline{Z} = \varnothing \in \mathfrak{F}$. Comme $Z \cup \varnothing = Z$ et $Z \cap \varnothing = \varnothing$, il vient que $\mathfrak{F} = \{Z, \varnothing\}$ est un corps booléen.

Exemple 5-3. Supposons que $\mathfrak{F}_0$ se compose de tous les événements élémentaires dans Z. Dans ce cas le corps booléen $\mathfrak{F}$ contient l'ensemble vide, tous les ensembles à un élément de Z, tous les ensembles à deux éléments de $Z, \ldots$ et, enfin, l'ensemble Z lui-même.

b) Notion de probabilité

En théorie des probabilités, on considère chaque épreuve liée à une expérience $z \in Z$ comme une variable aléatoire. Cela veut dire qu'on ne sait pas à l'avance laquelle des épreuves aura lieu à la suite

de l'expérience. Or, il est clair que, premièrement, chaque expérience donne lieu nécessairement à une certaine épreuve $z \in Z$ et, deuxièmement, jamais deux épreuves ne se produisent simultanément après une seule expérience.

Intuitivement, on conçoit que la *probabilité* de l'épreuve est une quantité mesurant la possibilité objective du cas donné (de l'épreuve donnée). Par là même, on associe à chacun des éléments $z \in Z$ un certain *poids* p, c.-à-d. un nombre réel positif assujetti à certaines restrictions, à savoir:

1) p est d'autant plus élevé qu'il y a plus de certitude quant à la réalisation de l'épreuve donnée $z \in Z$ au détriment des autres;

2) la somme des poids p de tous les cas possibles doit être égale à l'unité.

Si l'on associe à un certain élément $z \in Z$ le poids $p = 0$, cela veut dire que ce cas est impossible: il s'agit donc d'un *événement impossible*. Si un élément $z \in Z$ a le poids $p = 1$, le cas donné est un *événement certain*. En effet, en vertu de la restriction 2) tous les autres éléments de l'ensemble Z doivent avoir les poids $p = 0$, de sorte que les événements correspondant à ces éléments sont impossibles. Or, du moment qu'une épreuve doit se produire obligatoirement à la suite de l'expérience, ce sera justement celle qui a le poids $p = 1$.

La totalité des poids associés aux différents éléments de l'ensemble Z représente un certain ensemble P d'éléments p, tandis que le processus d'affectation des poids représente l'application de l'ensemble Z sur l'ensemble P; les probabilités $p \in P$ se définissent donc en fonction des épreuves $z \in Z$, ce qui peut être écrit sous la forme

$$p = p(z). \tag{5-3}$$

Conformément aux contraintes énoncées ci-dessus, les quantités $p(z)$ représentent des nombres réels vérifiant les conditions

$$p(z) \geqslant 0; \quad \sum_{z \in Z} p(z) = 1. \tag{5-4}$$

La totalité des poids $p(z)$ pour tous les $z \in Z$ constitue la *distribution des probabilités* sur l'espace des épreuves Z, et chaque poids $p \in P$ associé à une épreuve élémentaire $z \in Z$ est la *probabilité* de l'épreuve donnée.

c) Probabilité d'un événement aléatoire

Nous venons de définir l'événement aléatoire S comme un certain sous-ensemble de l'ensemble des épreuves Z liées à une expérience. Donc, l'événement aléatoire est l'ensemble réunissant certains élé-

ments de Z. Dans ce cas on entend par probabilité de l'événement S, désignée par P_S ou par $P(S)$, la somme des poids des éléments constitutifs de cet événement:

$$P(S) = P_S = \sum_{z \in S} p(z). \qquad (5\text{-}5)$$

Signalons deux cas particuliers de cette formule:

1) S ne contient aucun élément de Z, c.-à-d. qu'il est un ensemble vide; conformément à (5-5), la probabilité de l'ensemble vide est nulle:

$$P_{\varnothing} = 0; \qquad (5\text{-}6)$$

2) S se confond avec Z, c.-à-d. qu'il contient toutes les épreuves possibles. Comme aucune épreuve ne peut exister en dehors de Z, il est légitime de considérer l'espace Z dans sa totalité comme l'ensemble universel et lui associer le poids 1. Cela ne contredit nullement la formule (5-5), qui, considérée conjointement avec (5-4), donne:

$$P_Z = \sum_{z \in Z} p(z) = 1. \qquad (5\text{-}7)$$

d) Espace de probabilité

Les raisonnements précédents relatifs à la définition des probabilités ont un caractère quelque peu intuitif et ne garantissent pas la possibilité de calculer les probabilités des groupes d'événements arbitraires susceptibles de se produire dans la classe des événements $\mathfrak{F}$. Pour résoudre ce problème, il convient de considérer la distribution des probabilités non pas sur l'espace des épreuves Z mais sur la classe tout entière $\mathfrak{F}$ des événements liés à cet espace.

La *mesure de probabilité* sur une classe des événements $\mathfrak{F}$ représentant un corps booléen est la fonction réelle $P(S)$ vérifiant les conditions suivantes:

1) $P(S) \geqslant 0$ pour tout $S \in \mathfrak{F}$;

2) si $S_1, S_2, \ldots, S_n$ est la suite d'événements de $\mathfrak{F}$ incompatibles deux à deux, on a $P\left(\bigcup_{k=1}^{n} S_k\right) = \sum_{k=1}^{n} P(S_k)$;

3) $P(Z) = 1$.

Le couple $(\mathfrak{F}, P)$, c.-à-d. le corps d'ensembles et la mesure de probabilité définie sur lui, porte le nom de *champ de probabilité*.

Le triplet $(Z, \mathfrak{F}, P)$, c.-à-d. l'espace des événements élémentaires, le corps d'ensembles défini sur cet espace et la mesure de probabilité définie sur ce corps, porte le nom d'*espace probabilisé* ou d'*espace de probabilité*.

5-2. CALCUL DES PROBABILITÉS

a) Méthodes d'affectation de la mesure de probabilité

Il existe plusieurs méthodes pour associer des poids aux différents éléments z de l'espace Z. Le choix de la méthode à appliquer dans tel ou tel cas est déterminé par la nature de l'expérience envisagée et de l'information dont nous disposons. Considérons quelques méthodes principales.

1. *Détermination de la probabilité au moyen de la fréquence.* Supposons que l'expérience ayant l'espace des épreuves $Z = \{z_1, \ldots, z_m\}$ puisse être répétée un grand nombre de fois dans les mêmes conditions. Supposons qu'on ait procédé à N expériences et que le cas $z \in Z$ qui nous intéresse se soit produit N_z fois. Le nombre relatif de réalisations de z, c.-à-d. la quantité

$$q(z) = \frac{N_z}{N}, \tag{5-8}$$

est la *fréquence* de z. On s'assure sans peine que les fréquences $q(z)$ des épreuves possibles $z \in S$ vérifient les conditions (5-4).

Le nombre d'expériences n'étant pas élevé, la fréquence a un caractère manifestement aléatoire. C'est ainsi qu'en jetant une pièce de monnaie 10 fois, il arrive qu'on obtienne pile 2 fois, et aux 10 coups suivants, par exemple, 8 fois. Or, la pratique montre qu'avec l'accroissement du nombre des expériences la fréquence des cas isolés perd sensiblement son caractère aléatoire et tend (avec certaines oscillations) vers une certaine valeur moyenne, laquelle peut justement être assimilée à la probabilité de l'événement. Il importe de noter toutefois que cette tendance de la fréquence vers la probabilité, qu'on observe en augmentant le nombre des expériences, n'est pas la tendance vers une limite dans le sens mathématique.

Par exemple, il n'est pas impossible qu'en jetant la pièce de monnaie 10 fois on obtienne 10 fois pile. Du moment que le résultat de chaque coup ne dépend pas des résultats précédents, il n'est pas moins possible qu'on obtienne pile 1000 fois sur 1000 coups. Or, la probabilité d'un pareil événement est si infime que l'événement peut être considéré comme pratiquement irréalisable.

De façon générale, quand on multiplie les expériences, la fréquence tend vers la probabilité au point que la probabilité d'un écart quelque peu sensible de la fréquence par rapport à la probabilité devient négligeable. Donc, pourvu qu'on ait la possibilité de répéter l'expérience un grand nombre de fois dans les mêmes conditions, les fréquences des différentes épreuves $q(z)$ peuvent être retenues en qualité des probabilités correspondantes. Les questions relatives au nombre d'expériences que l'on peut considérer comme suffisant en déterminant la probabilité au moyen de la fréquence et au degré de

certitude des résultats obtenus par cette méthode seront traitées plus en détail dans les paragraphes consacrés aux méthodes de statistique mathématique.

Cependant, vu les difficultés liées aux répétitions multiples de l'expérience dans les conditions réelles, on est obligé d'avoir recours à d'autres méthodes de détermination de la probabilité.

2. Assez souvent (mais pas toujours) on peut utiliser la méthode basée sur le *principe d'égale probabilité.*

Ce principe entre en jeu lorsque nous n'avons aucune raison d'accorder la préférence à une épreuve au détriment des autres. On considère alors que toutes les épreuves sont également probables. Si l'espace des épreuves Z est constitué par m éléments, tandis que l'événement S contient r éléments de Z, la probabilité de cet événement sera

$$P(S)=\frac{r}{m}. \qquad (5\text{-}9)$$

3. La méthode la plus utilisée de détermination de la mesure de probabilité est la recherche des *probabilités a priori.* Ces probabilités sont déterminées en accumulant les données statistiques sur l'événement ou le phénomène en question pendant une période prolongée.

Généralement, dans diverses branches de la science, de la technique et de la vie sociale, on enregistre les données statistiques pour les événements ou phénomènes se rapportant à la catégorie des choses aléatoires, c.-à-d. accidentelles, dont les régularités ne se prêtent pas à la définition mathématique rigoureuse et dont l'apparition ne peut être prédite avec certitude. Apprenant la fréquence d'apparition de cet événement dans le passé, on arrive à établir sa probabilité et, partant, à prédire, avec un certain degré de certitude, l'apparition de cet événement à l'avenir.

4. Les probabilités a priori ne sont justifiables que si les conditions dans lesquelles l'événement s'est produit dans le passé existent au moment actuel. Or, bien souvent, ce n'est pas le cas. Aussi importe-t-il de préciser les conditions réelles existant au moment actuel; on y parvient en procédant à une expérience spéciale. Les probabilités recherchées tant à la base des données statistiques pour la période écoulée qu'à la base de l'expérience spéciale portent le nom de *probabilités a posteriori.*

b) Propriétés de la mesure de probabilité

Certaines propriétés de la mesure de probabilité s'avèrent fort utiles pour simplifier les calculs des probabilités. Examinons ces propriétés pour les événements X_1 et X_2 sur l'espace des épreuves Z.

9—0732

1. Si $X_1 \cap X_2 = \varnothing$, alors

$$P(X_1 \cup X_2) = P(X_1) + P(X_2), \tag{5-10}$$

ce qui découle directement de la définition de la mesure de probabilité.

2. A partir des conditions $X \cup \overline{X} = Z$ et $X \cap \overline{X} = \varnothing$, on trouve :

$$P(X \cup \overline{X}) = P(X) + P(\overline{X}) = 1, \tag{5-11}$$

d'où

$$P(\overline{X}) = 1 - P(X). \tag{5-12}$$

3. Soit $X_1 \subseteq X_2$. Alors X_2 peut être représenté sous la forme $X_2 = X_1 \cup (X_2 \setminus X_1)$, avec $X_1 \cap (X_2 \setminus X_1) = \varnothing$. Donc, $P(X_2) = P(X_1) + P(X_2 \setminus X_1)$. Puisque $X_2 \setminus X_1 \in Z$ et, par conséquent, $P(X_2 \setminus X_1) \geqslant 0$, on a :

$$P(X_2) \geqslant P(X_1); \tag{5-13}$$

$$P(X_2 \setminus X_1) = P(X_2) - P(X_1). \tag{5-14}$$

4. Pour X_1 et X_2 arbitraires appartenant à Z, représentons leur réunion comme la réunion de deux ensembles disjoints :

$$X_1 \cup X_2 = X_1 \cup [X_2 \setminus (X_1 \cap X_2)], \tag{5-15}$$

de sorte que

$$P(X_1 \cup X_2) = P(X_1) + P[X_2 \setminus (X_1 \cap X_2)]. \tag{5-16}$$

Puisque $X_1 \cap X_2 \subseteq X_2$, on a

$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2). \tag{5-17}$$

Exemple 5-4. On tire au hasard une carte d'un jeu de 52 cartes. Quelle est la probabilité de tirer :

1) un cœur ou le roi de trèfle?

2) un cœur ou un roi?

Désignons par X l'ensemble des cœurs, par Y l'ensemble des rois, par V le roi de trèfle, de sorte que $P(X) = 1/4$, $P(Y) = 1/13$, $P(V) = 1/52$.

1. Cherchons $P(X \cup V)$. Puisque $X \cap V = \varnothing$, on a $P(X \cup V) = P(X) + P(V) = \frac{1}{4} + \frac{1}{52} = \frac{7}{26}$.

2. Cherchons $P(X \cup Y)$. Puisque $X \cup Y \neq \varnothing$, on a $P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$. Or, $X \cap Y$ est le roi de cœur, donc, $P(X \cap Y) = 1/52$. Il vient

$$P(X \cup Y) = \frac{1}{4} + \frac{1}{13} - \frac{1}{52} = \frac{4}{13}.$$

5-3. PROBABILITÉS CONDITIONNELLES

a) Notion de probabilité conditionnelle

Considérons un événement T dans l'espace des épreuves Z. Par définition, la probabilité de cet événement est

$$P(T)=\sum_{z\in T} p(z). \tag{5-18}$$

Supposons maintenant qu'on a appris qu'un autre événement S vient de se produire. On demande de déterminer de quelle façon le fait de savoir la réalisation de S affectera-t-il la probabilité de T?

La probabilité d'un événement T soumise à la condition de réalisation d'un autre événement S porte le nom de *probabilité conditionnelle* de l'événement T et se désigne par $P_S(T)$ ou par $P(T \mid S)$. Au contraire, la probabilité $P(T)$ est la *probabilité absolue*, ou *inconditionnée*, de l'événement T. Il s'agit de rechercher la relation existant entre ces deux probabilités.

Prenons d'abord le cas élémentaire. Admettons que l'ensemble T ne comporte qu'un seul élement $z \in Z$. En déterminant la probabilité inconditionnée $p(z)$, nous supposons que n'importe quel $z \in Z$ peut se produire. Le nombre d'expériences étant suffisamment élevé, on a

$$p(z)=\frac{\text{Nombre d'expériences avec } z\in Z \text{ donné}}{\text{Nombre total d'expériences}}=\frac{N_z}{N}. \tag{5-19}$$

En imposant la condition de réalisation de l'événement S, on élimine automatiquement les épreuves ne confirmant pas la réalisation de S. Pour cette raison. en cherchant $p(z \mid S)$, on doit changer N en N_S, nombre d'expériences ayant amené S, et N_z en $N_{(z,\ S)}$, nombre d'expériences ayant amené tant z que S. Or, dans le cas considéré

$$(z,S)=z\cap S=\begin{cases} z & \text{si } z\in S; \\ \varnothing & \text{si } z\notin S. \end{cases} \tag{5-20}$$

Aussi $N_{(z,\ S)} = N_{z\cap S}$. De cette façon,

$$p(z\mid S)=\frac{N_{z\cap S}}{N_S}. \tag{5-21}$$

Divisant le numérateur et le dénominateur de cette expression par le nombre total d'expériences N, on obtient:

$$p(z\mid S)=\frac{p(z\cap S)}{p(S)}. \tag{5-22}$$

Passons maintenant au cas où T est un sous-ensemble arbitraire de l'ensemble Z. Se rappelant que la probabilité de l'événement est

égale à la somme des probabilités des éléments constitutifs de cet événement, on obtient:

$$P(T \mid S) = \sum_{z \in T} p(z \mid S) = \frac{\sum_{z \in T} p(z \cap S)}{P(S)}. \tag{5-23}$$

Examinons de plus près la somme figurant au numérateur. Elle est formée uniquement par les termes pour $z \in T$. D'autre part, on peut négliger dans cette somme les termes pour $z \notin S$, vu qu'on a alors $p(z \cap S) = 0$. La somme en question se compose donc des termes vérifiant les conditions $z \in T$ et $z \in S$. On peut donc sommer sur $z \in T \cap S$. Or, pour $z \in S$, on a $p(z \cap S) = p(z)$.

Aussi

$$\sum_{z \in T} p(z \cap S) = \sum_{z \in T \cap S} p(z) = P(T \cap S), \tag{5-24}$$

et l'on a:

$$P(T \mid S) = \frac{P(T \cap S)}{P(S)}, \tag{5-25}$$

ou, dans les notations plus usitées,

$$P(T \cap S) = P(T \mid S)\, P(S). \tag{5-26}$$

Parfois, le fait de savoir S réalisé n'affecte aucunément la probabilité de T. En d'autres mots, que S se soit produit ou non, la probabilité de T reste inchangée, de sorte que

$$P(T \mid S) = P(T). \tag{5-27}$$

On dit alors que T et S sont des événements *indépendants*. De pareils événements vérifient la relation

$$P(T \cap S) = P(T)\, P(S) \tag{5-28}$$

connue sous l'appellation de la *formule des probabilités composées*.

b) Variables aléatoires à deux dimensions

Très souvent, à la suite de l'expérience, nous cherchons à savoir non pas une mais plusieurs variables aléatoires, par exemple deux, x et y. C'est ainsi qu'en tirant une carte du jeu, ce sont à la fois sa couleur x et sa valeur y qui nous intéressent. Choisissant dans la forêt les arbres pour la construction, nous faisons attention à la hauteur de l'arbre x et à son diamètre y. Vérifiant le bon état d'un condensateur, nous établissons sa capacité x et la tension de claquage y.

Dans tous ces exemples on a affaire à deux ensembles de variables aléatoires $X = \{x_1, \ldots, x_n\}$ et $Y = \{y_1, \ldots, y_m\}$. Cependant ces ensembles ne sont pas indépendants mais liés entre eux d'une façon

déterminée: l'épreuve z d'une expérience unique donne lieu à deux variables aléatoires $x \in X$ et $y \in Y$, de sorte que $z = (x, y)$. Ainsi donc, l'ensemble des épreuves Z est le produit direct des ensembles X et Y:

$$Z = X \times Y. \tag{5-29}$$

Considérons quelques événements et leurs probabilités, que nous chercherons à savoir en examinant des variables aléatoires à deux dimensions. Pour illustrer les notions qui vont être introduites, nous reprendrons l'exemple de la hauteur x et du diamètre y des arbres dans la forêt.

Supposons qu'une expérience ait pour résultat deux épreuves $x \in X$ et $y \in Y$, par exemple $x = 30$ m et $y = 25$ cm. Les probabilités suivantes seront alors à considérer:

1) $p(x)$, probabilité pour que la hauteur de l'arbre soit de 30 m, abstraction faite de son diamètre;

2) $p(y)$, probabilité pour que le diamètre de l'arbre soit de 25 cm, abstraction faite de la hauteur de l'arbre. Les probabilités $p(x)$ et $p(y)$ sont les probabilités inconditionnées des paramètres x et y et fournissent les distributions des probabilités respectivement sur les ensembles X et Y de sorte que les conditions (5-4) soient vérifiées;

3) $p(x, y)$, la probabilité pour que l'arbre ait la hauteur de 30 m et le diamètre de 25 cm. C'est la distribution commune des probabilités de deux paramètres x et y définie sur l'espace $Z = X \times Y$ d'éléments $z = (x, y)$ et vérifiant les conditions suivantes:

$$p(x, y) \geqslant 0, \quad \sum_{x, y} p(x, y) = 1\,; \tag{5-30}$$

4) enfin, il se peut qu'on ait à considérer la probabilité conditionnée $p(x \mid y)$, c.-à-d. la probabilité pour que l'arbre de 25 cm de diamètre ait la hauteur de 30 m ; dans ce cas, on néglige simplement les arbres d'un diamètre autre que 25 cm.

Pour établir la relation entre les probabilités énumérées, désignons par x_i une épreuve $x \in X$ et par y_j une épreuve $y \in Y$ et considérons dans l'espace $Z = X \times Y$ deux événements $T_i = \{(x_i, y_1), \ldots, (x_i, y_m)\}$ et $S_j = \{(x_1, y_j), \ldots, (x_n, y_j)\}$.

L'événement T_i consiste dans la réalisation de l'épreuve donnée $x_i \in X$ et d'un quelconque $y \in Y$. Comme l'expérience donne toujours naissance à un certain $y \in Y$, l'événement T_i se réduit à la réalisation de x_i. On a donc $T_i = x_i$. De même $S_j = y_j$. Il est facile de voir par ailleurs que $T_i \cap S_j = (x_i, y_j)$. Par la formule (5-25), on trouve:

$$p(x_i \mid y_j) = \frac{p(x_i, y_j)}{p(y_j)}. \tag{5-31}$$

Du moment que x_i peut être n'importe quel $x \in X$ et que y_j peut être n'importe quel $y \in Y$, on a, en omettant les indices, pour tout

$x \in X$ et pour tout $y \in Y$:

$$p(x \mid y) = \frac{p(x, y)}{p(y)}, \tag{5-32}$$

ou, ce qui revient au même,

$$p(x, y) = p(x \mid y)\, p(y). \tag{5-33}$$

Si la probabilité de l'épreuve x ne dépend pas de la réalisation d'une épreuve concrète y, les épreuves x et y sont *indépendantes*. De telles épreuves vérifient la relation

$$p(x \mid y) = p(x), \tag{5-34}$$

de sorte que la formule (5-33) devient

$$p(x, y) = p(x)\, p(y). \tag{5-35}$$

Remarquons qu'en considérant une variable aléatoire à deux dimensions, aucune importance n'est attachée à ce que l'une des variables, x ou y, soit la première et l'autre la deuxième. Permutant x et y dans la formule (5-33), on a

$$p(x, y) = p(y \mid x)\, p(x). \tag{5-36}$$

Comparant (5-33) et (5-36), on constate que

$$p(x, y) = p(x)\, p(y \mid x) = p(y)\, p(x \mid y). \tag{5-37}$$

c) Formule de la probabilité totale

Considérons un espace d'épreuves Z. Soient $T \subseteq Z$ un événement dans l'espace Z, et le système d'ensembles $\{S_1, \ldots, S_l\}$ une partition de cet espace, de sorte que

$$Z = S_1 \cup \ldots \cup S_l, \quad S_i \cap S_k = \varnothing \text{ pour } i \neq k. \tag{5-38}$$

Assez souvent, il y a intérêt à définir la probabilité inconditionnée de l'événement T au moyen des probabilités conditionnelles $P(T \mid S_1), \ldots, P(T \mid S_l)$ et des probabilités inconditionnées $P(S_1), \ldots, P(S_l)$.

Pour avoir la relation cherchée, donnons à l'espace des épreuves Z la forme $Z = S_1 \cup \overline{S}_1$, où $\overline{S}_1 = S_2 \cup \ldots \cup S_l$. Alors

$$P(T) = P(T \mid S_1 \cup \overline{S}_1) = P[T \cap (S_1 \cup \overline{S}_1)] = \\ = P(T \cap S_1) + P(T \cap \overline{S}_1). \tag{5-39}$$

Considérons $\overline{S}_1$ comme un nouvel espace d'épreuves Z_1, c.-à-d. $Z_1 = \overline{S}_1 = S_2 \cup \ldots \cup S_l$ ou $\overline{S}_1 = S_2 \cup \overline{S}_2$, où $\overline{S}_2 = S_3 \cup \ldots \ldots \cup S_l$. Alors 5

$$P(T \cap \overline{S}_1) = P[T \cap (S_2 \cup \overline{S}_2)] = P(T \cap S_2) + P(T \cap \overline{S}_2). \tag{5-40}$$

Donc

$$P(T) = P(T \cap S_1) + P(T \cap S_2) + P(T \cap \bar{S}_2). \qquad (5\text{-}41)$$

Poursuivant les transformations analogues, on aboutit à

$$P(T) = \sum_i P(T \cap S_i) = \sum_i P(T \mid S_i)\, P(S_i), \qquad (5\text{-}42)$$

formule connue sous le nom de *formule de la probabilité totale.*

5-4. VARIABLES ALÉATOIRES CONTINUES ET LEURS DISTRIBUTIONS

a) Notion de variable aléatoire continue

Nous avons supposé jusqu'ici que l'espace des épreuves Z e·t un ensemble fini. Or, dans bien des cas, l'épreuve z liée à une expérience est une valeur quelconque se situant dans un certain intervalle $a \leqslant z \leqslant b$. Dans ce cas z est une variable aléatoire continue dont l'espace des épreuves se confond avec toute la gamme de ses valeurs possibles:

$$Z = \{z : a \leqslant z \leqslant b\}, \qquad (5\text{-}43)$$

contenant une infinité de points. On ne peut plus parler alors de la probabilité d'une épreuve isolée, car, le nombre des épreuves étant infini, le poids de chaque épreuve sera égal à zéro. Aussi la distribution des probabilités pour une variable aléatoire continue diffère-t-elle de la distribution pour le cas discret. On a deux modes de distribution, connus sous le nom de *fonction de répartition* et de *densité de probabilité.*

b) Fonction de répartition

Soient Z l'espace des épreuves pour une variable aléatoire à une dimension admettant des valeurs réelles, et R l'ensemble des nombres réels, dont les éléments seront désignés par x. La *fonction de répartition* des probabilités $F(x)$ définit la probabilité pour que la valeur de la variable aléatoire z ne soit pas supérieure à un x donné :

$$F(x) = P(-\infty < z \leqslant x). \qquad (5\text{-}44)$$

On arrive à se faire une idée générale de la fonction $F(x)$ en se basant sur quelques propriétés qu'elle possède:

1. Puisque pour $x_2 > x_1$ l'intervalle $(-\infty, x_2]$ est contenu tout entier à l'intérieur de l'intervalle $(-\infty, x_1]$, on a en vertu de (5-44) $F(x_2) \geqslant F(x_1)$. Par conséquent, $F(x)$ est une fonction non décroissante.

2. $F(-\infty) = 0$, car l'épreuve $z \in (-\infty, x]$ s'avère impossible quand $x \to -\infty$.

3. $F(+\infty) = 1$, car l'épreuve $z \in (-\infty, x]$ devient certaine quand $x \to +\infty$.

L'allure générale de la fonction vérifiant les propriétés définies est donnée sur la figure 5-2.

c) Densité de probabilité

Connaissant la fonction de répartition, on calcule la probabilité pour que la valeur de la variable aléatoire soit comprise dans un intervalle réduit, entre x et $x + \Delta x$. En vertu de (5-44), cette probabilité est égale à

$$F(x+\Delta x) - F(x) = \frac{F(x+\Delta x) - F(x)}{\Delta x} \Delta x. \qquad (5\text{-}45)$$

Le premier facteur du second membre de cette expression est la valeur de la probabilité par unité de longueur Δx de l'intervalle. Si,

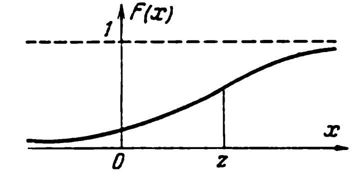


Fig. 5-2. Fonction de répartition d'une variable aléatoire continue

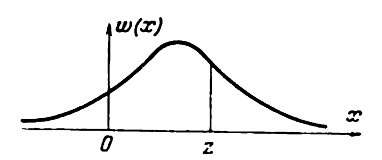


Fig. 5-3. Graphique de la densité de probabilité

pour $\Delta x \to 0$, la limite de cette expression existe, ce qui est généralement le cas pour la plupart des variables aléatoires continues, cette limite se désigne par $w(x)$ et reçoit l'appellation de *densité de probabilité*. De cette façon,

$$w(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = \frac{dF(x)}{dx}, \qquad (5\text{-}46)$$

de sorte que

$$F(x + \Delta x) - F(x) = w(x)\, \Delta x. \qquad (5\text{-}47)$$

On voit sur la figure 5-3 la forme générale du graphique de la fonction $w(x)$.

Mettons au clair quelques propriétés de la fonction $w(x)$. Soient a et b des points arbitraires de l'axe réel tels que $b > a$. Cherchons la signification de l'intégrale $\int_a^b w(x)\, dx$. Substituant tout formelle-

ment $dF(x)$ à $w(x)\,dx$, on obtient:

$$\int_a^b w(x)\,dx = \int_{F(a)}^{F(b)} dF(x) = F(b) - F(a). \qquad (5\text{-}48)$$

Or, en vertu de (5-44,) cette expression est la probabilité pour que la variable aléatoire z prenne une valeur située dans l'intervalle (a, b). Ainsi donc,

$$P(a < z < b) = \int_a^b w(x)\,dx. \qquad (5\text{-}49)$$

Cette formule a un cas particulier

$$\int_{-\infty}^{+\infty} w(x)\,dx = 1, \qquad (5\text{-}50)$$

montrant que l'aire limitée par la courbe de densité $w(x)$ et l'axe des abscisses est toujours égale à l'unité.

La formule (5-49) permet d'établir la liaison entre les fonctions $F(x)$ et $w(x)$ sous une forme intégrale. Comparant (5-44) avec (5-49), on trouve:

$$F(x) = \int_{-\infty}^{x} w(\xi)\,d\xi. \qquad (5\text{-}51)$$

La notion de fonction de répartition et celle de densité de probabilité s'étendent aisément au cas d'une variable aléatoire multidimensionnelle [33, 35]. Sans nous arrêter sur cette question, passons à l'examen de quelques lois de distribution des variables aléatoires continues.

d) Distribution uniforme

Si une variable aléatoire peut prendre avec probabilité égale des valeurs quelconques dans un intervalle de l'axe réel entre α et β et n'en peut prendre aucune qui soit en dehors de cet intervalle, c.-à-d. possède la densité de probabilité

$$w(z) = \begin{cases} \dfrac{1}{\beta - \alpha}, & \alpha \leqslant z \leqslant \beta; \\ 0, & z < \alpha, \quad z > \beta, \end{cases} \qquad (5\text{-}52)$$

on dit de cette variable qu'elle est *distribuée uniformément* dans l'intervalle $[\alpha, \beta]$.

Il est commode de caractériser la distribution uniforme par les paramètres $\nu = (\alpha + \beta)/2$ et $\omega = \beta - \alpha$, appelés *valeur moyenne*

et *étendue* de la distribution, et de la désigner donc par $R(\nu, \omega)$. On exprime sans peine la densité de probabilité $w(z)$ par les paramètres de la distribution $R(\nu, \omega)$ sous la forme

$$w(z)=\begin{cases}\frac{1}{\omega}, & \nu-\frac{\omega}{2}\leqslant z\leqslant\nu+\frac{\omega}{2};\\ 0, & z<\nu-\frac{\omega}{2}, \quad z>\nu+\frac{\omega}{2}.\end{cases} \tag{5-53}$$

Exemple 5-5. Quand on arrondit des nombres quelconques aux valeurs entières, les erreurs d'arrondissement peuvent prendre avec probabilité égale des valeurs de —0,5 à +0,5, c.-à-d. qu'elles obéissent à la distribution $R(0, 1)$.

Exemple 5-6. Quand on fait des lectures sur une échelle graduée en δ, les erreurs de lecture seront des variables aléatoires à distribution $R(0, \delta)$.

Exemple 5-7. Si le trafic du trolleybus s'effectue avec des intervalles de 10 mn, le temps d'attente pour un passager ignorant l'horaire du trafic sera une variable aléatoire caractérisée par la distribution $R(5, 10)$.

e) Distribution normale

La loi la plus importante de distribution d'une variable aléatoire est la *loi normale*, ou la loi de Gauss. La plus fréquente dans la pratique, elle est aussi la loi limite vers laquelle tendent de nombreuses autres lois de distribution jouant dans des conditions fort typiques.

La loi normale se caractérise par la fonction de densité de la forme

$$w(z)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(z-\nu)^2}{2\sigma^2}} \tag{5-54}$$

pour $-\infty<z<+\infty$. La distribution normale se définit par les paramètres ν et σ^2, dits *moyenne* et *variance*; elle sera désignée par

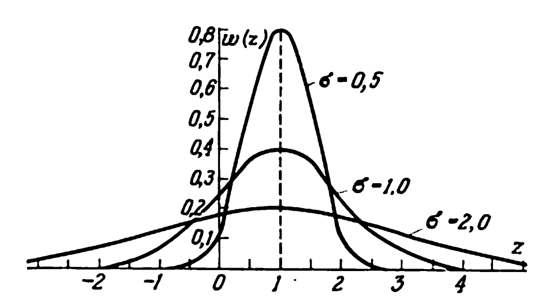


Fig. 5-4. Graphique de $w(z)$ pour la distribution normale

la suite $N(\nu, \sigma^2)$. On voit sur la figure 5-4 le graphique de la fonction (5-54) pour $\nu = 1,0$ et pour quelques valeurs de σ.

En théorie des probabilités, on attache une importance considérable à savoir la probabilité pour que la variable aléatoire tombe sur

un intervalle donné de l'axe réel (a, b). Dans le cas de la distribution normale cette probabilité est

$$P(a<z<b)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_a^b e^{-\frac{(z-\nu)^2}{2\sigma^2}}dz. \qquad (5\text{-}55)$$

On met cette expression sous une forme plus commode par changement de variable, en désignant $(z-\nu)/\sigma=t$. On a alors $dt = dz/\sigma$ et

$$P(a<z<b)=\frac{1}{\sqrt{2\pi}}\int_{\frac{a-\nu}{\sigma}}^{\frac{b-\nu}{\sigma}} e^{-\frac{1}{2}t^2}dt. \qquad (5\text{-}56)$$

Or, l'intégrale de la forme $\int e^{-\frac{1}{2}t^2}dt$ ne peut être exprimée au moyen de fonctions élémentaires. Aussi, pour le calcul de (5-56), a-t-on recours aux tables d'une fonction spéciale

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{1}{2}t^2}dt, \qquad (5\text{-}57)$$

appelée *intégrale de probabilité*, qu'on trouve dans la plupart des ouvrages consacrés au calcul des probabilités, par exemple dans [36]. Compte tenu de (5-57), la probabilité pour que la valeur aléatoire se situe dans l'intervalle (a, b) s'exprime par

$$P(a<z<b)=\Phi\left(\frac{b-\nu}{\sigma}\right)-\Phi\left(\frac{a-\nu}{\sigma}\right). \qquad (5\text{-}58)$$

Effectuant les calculs d'après la formule (5-58), il est utile de se rappeler les propriétés suivantes de l'intégrale de probabilité:

$$\Phi(0)=0; \quad \Phi(+\infty)=\frac{1}{2}; \quad \Phi(-x)=-\Phi(x). \qquad (5\text{-}59)$$

5-5. CARACTÉRISTIQUES NUMÉRIQUES DES VARIABLES ALÉATOIRES

a) Notion de caractéristiques numériques

La fonction de répartition ou la densité de probabilité sont les caractéristiques les plus complètes des variables aléatoires. Or, il existe un grand nombre de problèmes pratiques où il s'avère difficile ou même impossible de définir la fonction de répartition de façon complète.

Par ailleurs, pour avoir la solution de nombreux problèmes, il suffit de connaître un petit nombre de paramètres caractérisant la

variable aléatoire sous tel ou tel point de vue. Les paramètres numériques les plus employés, qui ont reçu l'appellation de *caractéristiques numériques* ou de *moments* des variables aléatoires, sont la *moyenne* et la *variance* (ou écart quadratique moyen). Se déduisant sans peine des données expérimentales, elles aident à se faire une idée d'ensemble sur la nature de distribution de la variable aléatoire.

b) Valeur moyenne (espérance mathématique) d'une variable aléatoire

Dans de nombreux problèmes, la variable aléatoire est assimilée à une épreuve et s'exprime par un nombre réel. Voici quelques exemples de variables aléatoires admettant une valeur numérique:

l'intensité du courant consommé par les locataires d'une maison;

la tension électrique dans le réseau à un moment donné;

la température de l'air à un moment donné de la journée;

le nombre de points marqué par le tireur donné en tirant à la cible;

le nombre de particules cosmiques atteignant la surface terrestre à l'unité de temps;

le nombre d'acheteurs dans le magasin à un moment donné, et ainsi de suite.

Dans tous ces cas, une caractéristique numérique bien importante de la variable aléatoire est sa valeur moyenne (ou simplement sa moyenne), dite aussi espérance mathématique.

Supposons que les éléments d'un espace des épreuves $Z = \{z_1, \ldots, z_m\}$ admettent une appréciation numérique. Si, au bout de N expériences, l'épreuve Z_1 a eu lieu r_1 fois, ..., z_m a eu lieu r_m fois, de sorte que $r_1 + \ldots + r_m = N$, la valeur moyenne de la variable aléatoire z, désignée par $M(z)$ ou $\bar{z}$, est égale à

$$M(z) = \bar{z} = \frac{z_1 r_1 + \ldots + z_m r_m}{N} = \sum_{i=1}^{m} z_i \frac{r_i}{N}. \tag{5-60}$$

Si le nombre d'expériences N était suffisamment élevé, la relation $r_i/N = p_i$ peut être interprétée comme la probabilité de l'épreuve z_i. Omettant les indices, on écrit la formule de la moyenne comme suit:

$$M(z) = \bar{z} = \sum_{z \in Z} z p(z). \tag{5-61}$$

La moyenne $M(z)$, déterminée par la formule (5-61), est souvent désignée sous le terme d'*espérance mathématique* de la variable aléatoire z.

La formule (5-61) peut servir à la détermination de la moyenne dans le cas où l'espace des épreuves Z est un ensemble fini. D'autre part, si z est une variable aléatoire continue caractérisée par une den-

sité de probabilité $w(z)$, on peut admettre dans (5-61) que $p(z)$ est la probabilité pour que la valeur de la variable aléatoire soit contenue dans les limites de z à $z + \Delta z$, ce qui veut dire qu'on peut poser $p(z) = w(z)\,\Delta z$. Passant à la limite pour $\Delta z \to 0$ et remplaçant respectivement Δz par dz et la somme par l'intégrale, on obtient:

$$M(z) = \bar{z} = \int_{-\infty}^{+\infty} zw(z)\,dz. \tag{5-62}$$

c) Valeur moyenne de la fonction d'une variable aléatoire

Bien souvent, dans la pratique, on rencontre des problèmes où la variable aléatoire ne peut être exprimée par aucun nombre. Par exemple, la production d'une usine se divise en produits de bonne qualité et rebut. Quand on jette une pièce de monnaie, on a comme épreuve soit pile, soit face. Une carte tirée au hasard d'un jeu se caractérise par sa valeur et sa couleur. Une impulsion radar réfléchie porte l'information relative à l'existence ou à l'absence de l'objectif, et ainsi de suite. Dans tous ces exemples les variables aléatoires n'ont qu'une différence qualitative, et non pas quantitative. Or, même lorsqu'il s'agit de variables n'ayant qu'une différence qualitative, on arrive généralement à les munir de caractéristiques quantitatives.

Nous assimilons une variable aléatoire au résultat d'une expérience, ou à une épreuve, l'expérience en question étant ou bien organisée de façon artificielle, ou bien survenant à la suite d'un processus naturel. Rappelons-nous que toute expérience exige certains frais et qu'elle ne vise pas seulement à satisfaire notre curiosité mais plutôt à atteindre un but déterminé. Le cas qui se réalise à la suite de l'expérience sera considéré comme favorable ou défavorable, suivant que le but qu'on s'est assigné a été atteint ou non. Le fait de réaliser le but envisagé est un gain, ou un profit, et le fait contraire, une perte, ou un dommage. Le profit et le dommage peuvent être exprimés tous les deux par des nombres, par exemple par certaines sommes de roubles.

Ainsi donc, il est légitime d'associer à toute épreuve $z \in Z$ une appréciation numérique, donc de définir l'application f de l'espace des épreuves Z sur l'ensemble des nombres réels R

$$f: Z \to R. \tag{5-63}$$

Cette application nous fournit une fonction réelle $f(z)$ définie sur Z, que l'on peut manipuler comme on ferait avec une variable aléatoire; proposons-nous de déterminer, en l'occurrence, sa valeur moyenne.

Soit $f(z)$ une fonction numérique définie sur un ensemble Z. Cela veut dire qu'aux éléments $z_1, \ldots, z_m$ de Z correspondent certaines valeurs de la fonction $f(z_1), \ldots, f(z_m)$. Ces valeurs ont les mêmes probabilités que les variables aléatoires $z_1, \ldots, z_m$. De cette façon, $p(z)$ représente la distribution des probabilités autant pour la variable aléatoire z que pour la fonction de la variable aléatoire $f(z)$. On peut donc rechercher la valeur moyenne de la fonction d'une variable aléatoire en employant les formules de la moyenne de la variable aléatoire par changement de (z) en $f(z)$, c.-à-d.

$$M[f(z)] = \overline{f(z)} \sum_{z \in Z} f(z)\, p(z) \qquad (5\text{-}64)$$

pour une variable aléatoire discrète et

$$M[f(z)] = \overline{f(z)} = \int_{-\infty}^{+\infty} f(z)\, w(z)\, dz \qquad (5\text{-}65)$$

pour une variable aléatoire continue.

Notons que la quantité $\overline{f(z)}$ calculée par (5-64) dépend de la distribution $p(z)$. Cela veut dire qu'une variation de la distribution des probabilités $p(z)$ sur l'espace Z affecterait la valeur moyenne de la fonction $\overline{f(z)}$. Mettant en valeur cette circonstance, on désigne quelquefois la valeur moyenne de la fonction $\overline{f(z)}$ par $f(p)$ en posant que

$$f(p) = \overline{f(z)}. \qquad (5\text{-}66)$$

L'emploi de la même lettre f pour la désignation de la valeur moyenne de la fonction et pour la désignation de la fonction elle-même ne donne lieu ici à aucun malentendu, car les fonctions $f(z)$ et $f(p)$ sont définies sur des ensembles différents: la fonction $f(z)$ l'est sur l'espace des épreuves Z, et la fonction $f(p)$, sur l'espace des distributions possibles des probabilités $p(z)$.

Exemple 5-8. L'usine produit des appareils. La probabilité pour qu'un appareil soit rebuté est p. Quel est le profit moyen par appareil réalisé si a est le prix de revient et b le prix réclamé de l'appareil?

Appliquant la formule (5-64), on trouve:

$$f(p) = (b - a)(1 - p) - ap = b(1 - p) - a.$$

Dans nombre de problèmes, on assimile la moyenne au gain moyen pour un grand nombre d'expériences. On admet dans pareils cas qu'il y a intérêt à procéder à l'expérience si $\overline{f(z)} = f(p) \geqslant 0$.

Une autre interprétation de la moyenne se présente quand on fait un pari. Supposons qu'un événement puisse se produire avec la probabilité p. Aux termes du pari, un parieur consent à payer une somme b si l'événement ne se produit pas, à condition que l'autre parieur paie une somme a si l'événement se produit. Ainsi donc, le

premier parieur a la probabilité p de toucher la somme a et la probabilité $1 - p$ de payer la somme b. Son gain moyen est

$$f(p) = ap - b(1 - p). \tag{5-67}$$

On admet que le pari est équitable si le gain moyen est nul, c.-à-d. si

$$\frac{a}{b} = \frac{1-p}{p}. \tag{5-68}$$

Théorème 5-1 (théorème de la moyenne). *Utilisant pour la définition de* $f(p)$ *l'expression* (5-64) *et remplaçant dans chaque terme* $f(z)$ *par* $\min\limits_z f(z)$, *on aboutit à la relation*

$$f(p) \geqslant \min_z f(z) \sum_{z \in Z} p(z) = \min_z f(z), \tag{5-69}$$

qui exprime le théorème de la moyenne.

d) Valeur moyenne de la fonction de deux variables aléatoires

Dans bien des cas on a affaire non pas à un seul mais à plusieurs événements aléatoires et l'on introduit alors une fonction réelle dépendant des épreuves qui se réalisent à la suite de chacun de ces événements. Dans une épreuve d'athlétisme, par exemple, le résultat de chaque participant est un événement aléatoire ; d'autre part, la place attribuée à l'équipe sera fonction tant des résultats des membres de cette équipe que des résultats des membres de l'équipe adverse.

Bornons-nous à considérer deux variables aléatoires indépendantes en désignant leurs espaces des épreuves respectivement par X et Y. Désignons par $p(x)$ et $q(y)$ les distributions des probabilités sur les espaces X et Y.

Soit $f(x, y)$ une fonction réelle définie pour tous $x \in X$ et $y \in Y$. On dit dans ce cas que la fonction $f(x, y)$ est définie sur le produit direct des ensembles $X \times Y$. La moyenne de cette fonction dépendra de la nature de la distribution des probabilités $p(x)$ et $q(y)$, de sorte que

$$M[f(x, y)] = \overline{f(x, y)} = f(p, q). \tag{5-70}$$

Pour définir $f(p, q)$, donnons-nous un $y \in Y$ déterminé. Quand y est donné, la fonction $f(x, y)$ n'est fonction que de x, et sa valeur moyenne, qui ne dépend que de $p(x)$, peut être désignée par $f_p(y)$ et se déduira de la formule (5-64) :

$$M_x[f(x, y)] = f_p(y) = \sum_x f(x, y)\, p(x). \tag{5-71}$$

La quantité $f_p(y)$ ne dépend plus de x mais seulement de y. En prenant sa moyenne suivant y on obtient précisément $f(p, q)$:

$$f(p, q) = M_y[f_p(y)] = \sum_y f_p(y)\, q(y) = \sum_{x,\, y} f(x, y)\, p(x)\, q(y). \qquad (5\text{-}72)$$

e) Espérance mathématique conditionnelle

Soient Z l'ensemble des épreuves et $f(z)$ une fonction numérique définie sur Z et telle que sa moyenne (ou espérance mathématique) se définisse par (5-64).

Supposons maintenant qu'un événement $S \subseteq Z$ se soit produit. A ce moment-là la distribution des probabilités n'est plus $p(z)$ mais $p(z \mid S)$: c'est la distribution conditionnelle. Substituant dans la formule de la moyenne $p(z \mid S)$ à $p(z)$, on aboutit à l'*espérance mathématique conditionnelle* de la fonction $f(z)$:

$$M(f \mid S) = \sum_{z \in Z} f(z)\, p(z \mid S) = \frac{1}{P(S)} \sum_{z \in Z} f(z)\, p(z \cap S) =$$

$$= \frac{\sum_{z \in S} f(z)\, p(z)}{\sum_{z \in S} p(z)}. \qquad (5\text{-}73)$$

f) Propriétés de la valeur moyenne

1. La valeur moyenne d'une variable non aléatoire est égale à la variable elle-même.

Soit c une variable non aléatoire; il s'agit donc, en réalité, non pas d'une variable mais d'une constante. On peut admettre dans ce cas que c est l'épreuve dont l'espace Z n'a pour éléments que c seul, de sorte que $p(c) = 1$. On a par la formule (5-61):

$$M(c) = cp(c) = c. \qquad (5\text{-}74)$$

2. Le facteur constant peut être sorti de sous le signe de la moyenne:

$$M(cz) = \sum_{z \in Z} czp(z) = c \sum_{z \in Z} zp(z) = cM(z). \qquad (5\text{-}75)$$

3. La moyenne de la somme de deux variables aléatoires est la somme de leurs moyennes.

Soient x et y deux variables aléatoires distribuées sur les espaces des épreuves X et Y avec les probabilités $p(x)$ et $q(y)$. Notons $f(x, y) = x+y$. Par la formule (5-72), on a:

$$\overline{x+y} = \sum_{x,\, y} (x+y)\, p(x)\, q(y) = \sum_x xp(x) \sum_y q(y) + \sum_y yq(y) \sum_x p(x) =$$

$$= \sum_x xp(x) + \sum_y yq(y) = \bar{x} + \bar{y}. \qquad (5\text{-}76)$$

4. L'espérance mathématique d'une variable aléatoire centrée est égale à zéro.

Soient x une variable aléatoire et ν sa moyenne. La variable aléatoire $\overset{0}{z} = z - \nu$ est dite variable aléatoire centrée. Elle a pour expression

$$M(\overset{0}{z}) = M(z - \nu) = M(z) - \nu = 0. \tag{5-77}$$

g) Moments. Variance. Ecart quadratique moyen

Les caractéristiques numériques les plus importantes d'une variable aléatoire sont ses moments, qui se divisent en initiaux et centrés.

Le *moment initial* d'ordre s de la variable aléatoire z se définit par la formule

$$\alpha_s(z) = \sum_z z^s p(z) \tag{5-78}$$

pour une variable aléatoire discrète et par la formule

$$\alpha_s(z) = \int_{-\infty}^{+\infty} z^s w(z)\, dz \tag{5-79}$$

pour une variable aléatoire continue.

Il est facile de voir que la moyenne $M(z) = \overline{z}$, désignée plus loin par ν ou ν_z, représente le moment initial du premier ordre $\alpha_1(z)$. Le moment initial du second ordre est la moyenne du carré de la variable aléatoire:

$$\alpha_2(z) = \sum_z z^2 p(z) = M(z^2) = \overline{z^2}. \tag{5-80}$$

Dans le cas où l'on utilise en qualité de variable aléatoire une variable aléatoire centrée $\overset{0}{z} = z - \nu$, les formules (5-78) et (5-79) donnent les expressions pour les moments centrés d'ordre s, désignés par $\mu_s(z)$. On voit de la relation (5-77) que le moment centré du premier ordre est nul.

Une caractéristique numérique très importante d'une variable aléatoire est son moment centré du second ordre appelé *variance* de la variable aléatoire z et désigné par $D(z)$ ou D_z. Les formules définissant la variance sont

$$D(z) = \mu_2(z) = \sum_z (z - \nu_z)^2 p(z) \tag{5-81}$$

pour une variable aléatoire discrète et

$$D(z) = \mu_2(z) = \int_{-\infty}^{+\infty} (z - \nu_z)^2 w(z)\, dz \tag{5-82}$$

pour une variable aléatoire continue. La variance caractérise les écarts des différentes valeurs de la variable aléatoire par rapport à sa valeur moyenne, c.-à-d. qu'elle caractérise la dispersion de cette variable. Plus la variance est petite, plus les différentes valeurs de la variable aléatoire se concentrent autour de la moyenne.

Dans bien des cas, la variance s'avère d'un emploi peu commode dans la pratique, car sa dimension est celle du carré de la variable aléatoire. Aussi adopte-t-on souvent, comme caractéristique de la dispersion d'une variable aléatoire, la racine carrée de la variance, désignée sous le terme d'*écart quadratique moyen* et notée par σ ou σ_z:

$$\sigma_z = \sqrt{D(z)}. \tag{5-83}$$

Notons quelques propriétés de la variance.

1. La variance d'une variable non aléatoire (constante) est égale à zéro:

$$D(c) = M[(c - c)^2] = M(0) = 0. \tag{5-84}$$

2. Une constante peut être sortie de sous le signe de la variance en l'élevant à la puissance deux:

$$D(cz) = M[(cz - cv)^2] = c^2 D(z). \tag{5-85}$$

D'autres propriétés de la variance seront examinées plus tard.

h) Régression et corrélation

Lorsqu'il y a plusieurs variables aléatoires, par exemple deux, la moyenne et la variance peuvent caractériser chacune d'elles isolément. Or, en pareils cas, il importe beaucoup de connaître l'influence exercée par une de ces variables sur l'autre, c.-à-d. de tenir compte de la liaison réciproque établie entre les variables aléatoires. La régression et la corrélation sont appelées justement à fournir l'expression quantitative de cette liaison.

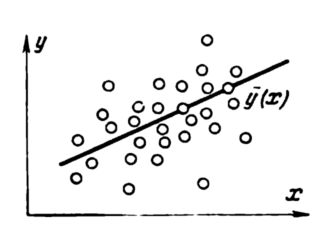


Fig. 5-5. Ligne de régression $\bar{y}(x)$

Soient x et y deux variables aléatoires, analogues à celles qu'on a examinées au § 5-3. Pour fixer les idées, supposons que x est la hauteur et y le diamètre des arbres dans un compartiment de la forêt. A chaque arbre correspondra un point sur le plan (x, y), tandis que la totalité des arbres aura pour expression l'ensemble des points de la figure 5-5. Dans le cas examiné les grandeurs $v_x = \bar{x}$ et $v_y = \bar{y}$ fournissent les valeurs moyennes de la hauteur et du diamètre des arbres, et σ_x et σ_y caractérisent la dispersion de la hauteur et du diamètre par rapport à leurs valeurs moyennes respectives.

En plus de la hauteur moyenne et du diamètre moyen, on attache un intérêt considérable à l'étude de la variation du diamètre en fonction de la hauteur de l'arbre. Or, pour les arbres de même hauteur x, le diamètre y est une variable aléatoire. Aussi ne parlera-t-on que de la valeur moyenne du diamètre $\bar{y}$ en fonction de la hauteur x, c.-à-d. qu'on se bornera à rechercher la seule valeur $\bar{y}(x)$ qui est la moyenne conditionnelle $M(y \mid x)$. Utilisant l'expression (5-73) et désignant par $p(x, y)$ la probabilité commune des valeurs données de x et de y, on trouve:

$$\bar{y}(x) = M(y \mid x) = \frac{\sum_y y p(x, y)}{\sum_y p(x, y)}. \tag{5-86}$$

Déterminant $\bar{y}(x)$ pour différents x, on arrive à construire une ligne traduisant graphiquement cette fonction (voir fig. 5-5) et appelée *ligne de régression* de y par rapport à x. On obtient par analogie la fonction $\bar{x}(y) = M(x \mid y)$, dite régression de x par rapport à y.

Etudiant la régression, nous ne prendrons que le cas le plus simple et le plus fréquent : il s'agit de la régression linéaire, avec la ligne de régression se présentant sous la forme d'une droite d'équation

$$\bar{y}(x) = a + b(x - \bar{x}). \tag{5-87}$$

Les coefficients a et b seront choisis de façon à concentrer au maximum les points (x, y) dans le voisinage de la droite $\bar{y}(x)$, ce qu'on réalise en imposant la condition

$$\psi(a, b) = M\{[y - \bar{y}(x)]^2\} = \min. \tag{5-88}$$

La condition (5-88), conjointement avec (5-87), fournit le système d'équations suivant pour la détermination de a et de b:

$$\left.\begin{aligned} \bar{y} - a = 0; \\ M[y(x - \bar{x})] - b\sigma_x^2 = 0. \end{aligned}\right\} \tag{5-89}$$

Appelons *covariance* entre x et y la quantité

$$\mu_{xy} = \text{cov}(x, y) = M[(y - \bar{y})(x - \bar{x})]. \tag{5-90}$$

Il est facile de voir que μ_{xy} peut être représentée sous la forme

$$\mu_{xy} = M[y(x - \bar{x}) - \bar{y}(x - \bar{x})] =$$

$$= M[y(x - \bar{x})] - \bar{y}M(x - \bar{x}) = M[y(x - \bar{x})]. \tag{5-91}$$

Compte tenu de (5-91), on tire de (5-89) les valeurs de a et de b définissant les lignes de régression:

$$a = \bar{y}; \qquad b = \frac{\mu_{xy}}{\sigma_x^2}. \tag{5-92}$$

La covariance μ_{xy} est susceptible de servir de mesure de liaison réciproque existant entre deux variables aléatoires x et y. Or, du moment qu'on s'est borné au seul cas où la ligne de régression est une droite, la covariance ne caractérisera pas une liaison réciproque de nature quelconque mais seulement la liaison linéaire, aux termes de laquelle une variable aléatoire tend à croître ou à diminuer en moyenne suivant la loi linéaire avec l'accroissement ou la diminution d'une autre variable aléatoire. Pour mieux représenter le changement de la covariance en fonction de la liaison entre deux variables aléatoires, il est utile d'examiner les cas extrêmes.

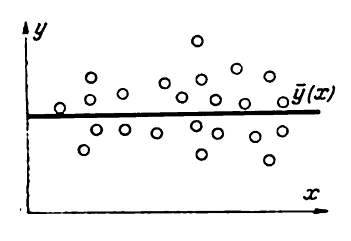


Fig. 5-6. Ligne de régression pour des variables aléatoires indépendantes

1. $\mu_{xy} = 0$. Cela signifie que $b = 0$, $\bar{y} = a$ et la ligne de régression est horizontale, ainsi qu'il est montré sur la figure 5-6. On voit que, dans les points expérimentaux, les valeurs de y se groupent autour de la valeur de a, quelle que soit la valeur de x. Donc, dans le cas donné, y et x sont des variables aléatoires indépendantes.

2. La liaison entre x et y sera plus prononcée quand ces deux variables seront liées par une dépendance fonctionnelle linéaire. Dans ce cas tous les points expérimentaux tomberont sur la ligne de régression, de sorte que $y = \bar{y}(x)$. Remplaçant $\bar{y}(x)$ par y, on récrit l'équation de la ligne de régression sous la forme

$$y-\bar{y} = \frac{\mu_{xy}}{\sigma_x^2}(x-\bar{x}). \tag{5-93}$$

Cherchons la variance σ_y^2 pour le cas donné:

$$\sigma_y^2 = M[(y-\bar{y})^2] = \frac{\mu_{xy}^2}{\sigma_x^2}, \tag{5-94}$$

d'où

$$\mu_{xy} = \sigma_x\sigma_y. \tag{5-95}$$

De cette façon, suivant que la liaison entre x et y est plus ou moins prononcée, la covariance μ_{xy} change entre 0 et $\sigma_x\sigma_y$.

La covariance μ_{xy} ne convient pas très bien pour apprécier le degré de liaison entre les variables x et y, du moment qu'elle dépend des variances des variables aléatoires elles-mêmes. A cette fin, on emploie plutôt le *coefficient de corrélation*

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x\sigma_y}, \tag{5-96}$$

qui varie de 0 (lorsque les variables aléatoires sont indépendantes) jusqu'à 1 (les variables aléatoires étant linéairement dépendantes).

La notion de covariance permet d'écrire des relations simples pour la moyenne du produit et la variance de la somme de deux variables aléatoires. Nous laissons au lecteur le soin de vérifier que, dans les cas indiqués, on a les relations suivantes:

$$\overline{xy} = \bar{x}\bar{y} + \mu_{xy}; \tag{5-97}$$

$$D(x+y) = D(x) + D(y) + 2\mu_{xy}. \tag{5-98}$$

Quand les variables aléatoires x et y sont indépendantes, on a

$$\overline{xy} = \bar{x}\,\bar{y}; \tag{5-99}$$

$$D(x+y) = D(x) + D(y). \tag{5-100}$$

5-6. PROCESSUS STOCHASTIQUES DISCRETS

a) Catégories de processus stochastiques discrets

Supposons qu'un système donné, soumis à l'influence de l'ensemble des conditions extérieures, puisse changer d'état, de façon aléatoire, à des instants discrets notés conventionnellement 0, 1, ..., les intervalles entre ces instants étant désignés sous le terme de *pas* ou *étapes*. Ce système, à l'étape n de son évolution, prend l'état x_n que l'on peut considérer comme résultat d'une expérience, ou épreuve, et qui constitue donc un élément de l'ensemble des épreuves possibles à l'étape donnée:

$$Z_n = \{z_0^n,\ z_1^n,\ \ldots,\ z_{L_n}^n\}.$$

Nous admettons que l'ensemble Z_0 soit constitué d'un seul élément z_0^0, représentant l'état initial du système. La suite des états pris par le système x_0, x_1, . . . est appelée *processus stochastique* [7].

Pour que le processus stochastique soit défini, il faut qu'on connaisse à chaque étape la distribution des probabilités sur l'espace des épreuves, définissant les états que le système peut prendre à l'étape suivante. Dans le cas général, cette distribution des probabilités dépend des états parcourus à toutes les étapes précédentes et se définit, pour tout n, par la collection des probabilités conditionnelles

$$p\,(x_{n+1} = z_k^{n+1} \mid x_0, \ldots, x_n), \qquad k = 0, 1, \ldots, L_n \tag{5-101}$$

avec

$$\sum_{k=0}^{L_n} p\,(x_{n+1} = z_k^{n+1} \mid x_0, \ldots, x_n) = 1. \tag{5-102}$$

Le processus sera poursuivi à l'infini, à moins qu'il ne soit interrompu à dessein à une étape quelconque. En pratique, nous sommes obligés toujours d'interrompre le processus après un certain nombre fini d'étapes, c.-à-d. d'avoir affaire aux processus dits *finis* ou

multiétapes. Or, les processus infinis sont d'utiles appproximations des processus finis, plus réels, d'autant plus qu'ils s'avèrent bien souvent plus simples du point de vue analytique.

Un processus stochastique dont nous ne pouvons modifier le déroulement ni, par conséquent, la distribution des probabilitités (5-101) est dit *incontrôlable.*

En fonction du type de la distribution des probabilités (5-101), on distingue plusieurs catégories de processus stochastiques, dont les plus répandus sont les processus à valeurs indépendantes, les processus d'essais indépendants et les chaînes markoviennes.

Un processus stochastique fini est dit processus à valeurs indépendantes si la distribution des probabilités sur l'espace des épreuves à chaque étape ne dépend pas des étapes précédentes:

$$p\,(x_n = z_k^n \mid x_0, \ldots, x_{n-1}) = p\,(x_n = z_k^n). \qquad (5\text{-}103)$$

Il est quelquefois plus commode de désigner cette distribution des probabilités simplement par $p_n(z)$, où $z \in Z_n$, $n = 0, 1, \ldots$

Une propriété fort importante d'un processus à valeurs indépendantes consiste en ce que toute suite d'états $x_0, x_1, \ldots, x_n$ peut être considérée comme un ensemble d'épreuves indépendantes. Par conséquent, la probabilité de réalisation d'une suite pareille est égale au produit des probabilités des épreuves constituant cette suite, ce qui peut être écrit sous la forme

$$p\,(x_0, x_1, \ldots, x_n) = p_0\,(x_0)\, p_1\,(x_1) \ldots p_n\,(x_n),$$
$$x_i \in Z_i, \quad i = 0, 1, \ldots, n. \qquad (5\text{-}104).$$

Il existe un cas particulier important des processus à valeurs indépendantes: ce sont les processus d'essais indépendants. Le *processus d'essais indépendants* est un processus à valeurs indépendantes qui a, à chaque étape, le même espace des épreuves $Z = \{z_0, z_1, \ldots, z_L\}$ et la même distribution des probabilités sur cet espace:

$$p_n\,(z) = p_m\,(z) = p\,(z) \qquad (5\text{-}105)$$

pour tout $z \in Z$ et pour tous m et n. En d'autres mots, le processus d'essais indépendants n'est qu'une répétition multiple d'un même essai dans les mêmes conditions.

De même que pour les processus à valeurs indépendantes, la probabilité d'obtenir une suite concrète d'épreuves $x_0, x_1, \ldots, x_n$ au cours des essais indépendants est égale au produit des probabilités d'obtention des épreuves isolées.

Le processus stochastique reçoit l'appellation de *chaîne markovienne* si l'espace des états $Z = \{z_0, z_1, \ldots, z_L\}$ est le même pour toutes les étapes et que la distribution des probabilités des épreuves à chaque étape ne dépende que des épreuves qui se sont réalisées

à l'étape précédente, de sorte que

$$p(x_{n+1}=z \mid x_0, \ldots, x_n) = p(x_{n+1}=z \mid x_n). \qquad (5\text{-}106)$$

Il est plus commode d'introduire pour les probabilités de la forme (5-106) des notations quelque peu différentes. Supposons qu'à la n-ième étape le système ait pris l'état $z_i \in Z$. Alors la distribution des probabilités (5-106) définira la probabilité pour que le système se trouve à la $(n+1)$-ième étape dans l'état $z_j \in Z$. Dans le cas général, cette probabilité dépend de l'indice de l'étape et peut être désignée par $p_{ij}(n)$. De cette façon, les probabilités

$$p_{ij}(n) = p(x_{n+1}=z_j \mid x_n = z_i) \qquad (5\text{-}107)$$

vérifiant la condition

$$p_{ij}(n) \geqslant 0, \quad \sum_{j=0}^{L} p_{ij}(n) = 1 \qquad (5\text{-}108)$$

sont les probabilités de passage à la n-ième étape de l'état z_i à l'état z_j.

Dans le cas où les probabilités de passage $p_{ij}(n)$ dépendent de l'indice de la n-ième étape, la chaîne markovienne est dite *non homogène*. Or ce sont les chaînes markoviennes *homogènes* qui jouent un rôle très important dans différentes applications ; dans ces chaînes les probabilités de passage restent les mêmes pour toute étape et peuvent être désignées simplement par p_{ij}.

b) Processus d'essais indépendants à deux épreuves. Distribution binomiale des probabilités

Considérons de façon plus approfondie les cas où, dans un processus d'essais indépendants, il n'y a que deux épreuves pour chaque essai qui nous intéressent, à savoir : un certain événement S se produira-t-il ou ne se produira-t-il pas à la suite de cet essai ? Une des épreuves sera appelée cas favorable ou *succès*, et l'autre, cas défavorable ou *insuccès*.

Désignons par p la probabilité du succès d'un essai isolé, et par $q = 1 - p$, la probabilité de l'insuccès. Proposons-nous de rechercher la probabilité $w(r, n, p)$ d'avoir r succès au cours de n essais.

Pour plus de commodité, désignons le succès par 1 et l'insuccès par 0. Alors à la réalisation de r succès au cours de n essais correspondra une suite de n épreuves constituée de r unités et de $n - r$ zéros. Du moment que les épreuves isolées sont indépendantes, la probabilité de cette suite, égale au produit des probabilités des épreuves isolées, aura pour expression

$$p^r q^{n-r}. \qquad (5\text{-}109)$$

Le nombre total des différentes suites de n éléments contenant r unités est égal au nombre des combinaisons de n éléments r à r,

c.-à-d.

$$\binom{n}{r}=\frac{n!}{r!\,(n-r)!}. \tag{5-110}$$

Puisque ces suites sont toutes incompatibles (la réalisation de l'une d'elles rend impossible la réalisation de toute autre), la probabilité d'apparition de l'une quelconque d'entre elles sera égale à

$$w(r, n, p)=\binom{n}{r}p^r q^{n-r}. \tag{5-111}$$

Il est facile de vérifier que la probabilité totale de réalisation d'un nombre quelconque d'événements de 0 à n est égale à l'unité.

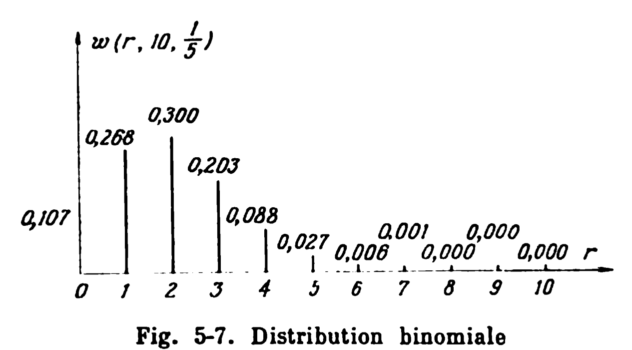


Fig. 5-7. Distribution binomiale

La distribution des probabilités définie par (5-111) porte le nom de *distribution binomiale*. On se rend comte sans peine de la nature de la distribution binomiale en regardant à la figure 5-7 le graphique des ordonnées de cette distribution en fonction du nombre r pour $n = 10$ et $p = 1/5$, c.-à-d. le graphique de distribution $w(r, 10, 1/5)$.

On s'attache très souvent à savoir la valeur de r telle que la distribution binomiale passe par le maximum. Calculons d'abord le rapport entre deux ordonnées voisines de la distribution binomiale

$$\frac{w(r, n, p)}{w(r-1, n, p)}=\frac{\binom{n}{r}p^r q^{n-r}}{\binom{n}{r-1}p^{r-1}q^{n-r+1}}=\frac{(n-r+1)p}{rq}. \tag{5-112}$$

Ce rapport ne sera pas inférieur à l'unité tant qu'il y aura $r \leqslant$ $\leqslant (n+1)\,p$. Par conséquent, la fonction $w(r, n, p)$ passe par le maximum quand

$$r=\text{ent}\,(n+1)\,p, \tag{5-113}$$

ent x signifiant partie entière du nombre x.

La distribution binomiale a pour caractéristiques numériques sa moyenne ν_r et sa variance σ_r^2, qui sont définies par les relations

$$\nu_r = np\,; \qquad \sigma_r^2 = npq. \tag{5-114}$$

c) Distribution de Poisson

Il est extrêmement difficile de calculer les valeurs de $w(r, n, p)$ quand n sont grands. Le calcul de ces probabilités se trouve cependant grandement facilité par l'application de certaines méthodes approximatives, dont l'une consiste à substituer à la distribution binomiale la distribution de Poisson.

Soit $w(r, n, p)$ la distribution binomiale. On constate sans peine que la limite de cette distribution pour $n \to \infty$ et $p \to 0$ et pour $np = a = \text{const}$, de sorte que $p = a/n$, est égale à

$$\lim_{n\to\infty} w(r, n, n/a) = \frac{a^r e^{-a}}{r!}\,. \tag{5-115}$$

La distribution de la forme (5-115) se désigne par $w(r, a)$ et s'appelle *distribution de Poisson*. On a donc

$$w(r, a) = \frac{a^r e^{-a}}{r!}\,. \tag{5-116}$$

Une particularité caractéristique de la distribution de Poisson est que la moyenne et la variance de cette distribution sont les mêmes et sont égales à a.

La distribution de Poisson est une bonne approximation de la distribution binomiale pour p petits. Ainsi, pour $p \leqslant 0{,}01$ on peut remplacer le calcul de la distribution binomiale par le calcul de la distribution de Poisson en commençant par $n = 10$.

Bien qu'obtenue comme cas limite de la distribution binomiale, la distribution de Poisson fait partie d'une classe indépendante des processus stochastiques, dite *classe des phénomènes aléatoires rares*.

Considérons la suite d'événements S se succédant avec des intervalles de temps aléatoires. Le nombre de tels événements se réalisant dans un intervalle de durée τ sera une variable aléatoire ayant pour moyenne $\lambda\tau$, où λ est le nombre moyen d'événements à l'unité de temps. Proposons-nous de déterminer la probabilité de réalisation de r événements exactement dans l'intervalle τ.

Les événements S peuvent être considérés comme rares dans ce sens qu'on peut partager l'intervalle τ en de brefs intervalles partiels de durée $\Delta\tau$ dont chacun ne peut donner lieu qu'à la réalisation d'un seul événement S. Le nombre total de tels intervalles est $n = \tau/\Delta\tau$ et la probabilité pour que l'événement se réalise dans l'un des in-

tervalles $\Delta\tau$ est $p = \lambda\tau/n$, de sorte que $np = \lambda\tau$. Admettant que l'événement S se réalise dans chacun des intervalles $\Delta\tau$ sans liaison avec sa réalisation ou non-réalisation dans d'autres intervalles, on aboutit à un processus d'essais indépendants caractérisé par la distribution des probabilités $w(r, n, p)$, où $np = \lambda\tau$.

Or, l'hypothèse de l'unicité de l'événement S sur l'intervalle $\Delta\tau$ ne se justifie que lorsque la durée de cet intervalle est extrêmement courte, c.-à-d. à la limite, quand $\Delta\tau \to 0$. Puisqu'on a à ce moment $n \to \infty$ et $p \to 0$, avec $np = \lambda\tau = \text{const}$, on conçoit que la classe des phénomènes rares obéit à la loi de Poisson avec la distribution des probabilités

$$w(r, \lambda\tau) = \frac{(\lambda\tau)^r e^{-\lambda\tau}}{r!}. \tag{5-117}$$

d) Distribution exponentielle. Notion de fiabilité

Dans de nombreux problèmes, on s'attache à savoir la distribution de probabilité pour les intervalles de temps entre les événements aléatoires dans le processus d'essais indépendants. De la formule (5-117), pour $r = 0$ et $r = 1$, on tire:

$e^{-\lambda\tau}$, probabilité pour qu'aucun événement ne se réalise dans l'intervalle de durée τ;

$\lambda\tau e^{-\lambda\tau}$, probabilité pour qu'il se réalise exactement un événement dans l'intervalle de durée τ.

La probabilité pour qu'il se réalise plus d'un événement dans l'intervalle de durée τ est égale à

$$1-(e^{-\lambda\tau}+\lambda\tau e^{-\lambda\tau}) = 1-\left\{\left[1-\lambda\tau+\frac{(\lambda\tau)^2}{2!}-\dots\right]-\right.$$
$$\left.-\lambda\tau\left[1-\lambda\tau+\frac{(\lambda\tau)^2}{2!}-\dots\right]\right\} = \frac{(\lambda\tau)^2}{2}+\dots \tag{5-118}$$

On voit que cette probabilité est proportionnelle à la quantité $(\lambda\tau)^2$, ce qui signifie que pour des $\lambda\tau$ faibles la probabilité de réalisation de deux ou plusieurs événements dans l'intervalle τ est négligeable, tandis que la probabilité de réalisation d'un seul événement dans cet intervalle est

$$F(\lambda, \tau) = 1 - e^{-\lambda\tau}, \qquad \tau \geqslant 0. \tag{5-119}$$

Au fond, l'expression (5-119) définit la probabilité pour que le temps pendant lequel un événement se produit ne soit pas supérieur à τ. Pour obtenir la densité de cette probabilité, ou la probabilité pour que l'événement se produise à l'instant τ, dérivons (5-119) par rapport à τ:

$$w(\lambda, \tau) = \lambda e^{-\lambda\tau}, \qquad \tau \geqslant 0. \tag{5-120}$$

La densité de probabilité pour τ étant connue, on détermine aisément le temps moyen pendant lequel aucun événement n'a lieu:

$$\tau_{\text{moy}} = \int_0^\infty \tau w(\lambda\tau)\, d\tau = \frac{1}{\lambda}. \qquad (5\text{-}121)$$

Les formules (5-119) et (5-120) définissent la *distribution exponentielle*, qui joue un rôle important en théorie de la fiabilité.

Imaginons un dispositif formé d'un grand nombre d'éléments isolés, dont chacun peut tomber en panne avec le temps. Dans la plupart des cas les pannes des divers éléments se produisent sans liaisons entre elles, et leur suite peut être considérée comme un processus d'événements indépendants, dans lequel λ définit le nombre de pannes à l'unité de temps et s'appelle *intensité des pannes*, tandis que la quantité $1/\lambda$ définit, conformément à (5-121), le temps moyen de fonctionnement sans défaillance.

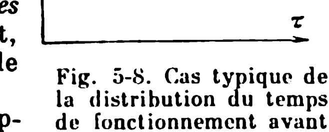


Fig. 5-8. Cas typique de la distribution du temps de fonctionnement avant la première panne

Pour caractériser la fiabilité d'un appareillage, on emploie habituellement la quantité $h(\tau)$, dite *risque de panne*, qui exprime la densité de probabilité de la panne à l'instant τ à condition qu'aucune panne n'ait eu lieu jusque-là.

Déterminons au préalable la quantité $h(\tau)\Delta\tau$ qui fournit la probabilité pour que la panne se produise pendant l'intervalle $\Delta\tau$, sachant qu'aucune panne ne s'est manifestée pendant la période τ précédente. D'après la formule de la probabilité conditionnelle

$$h(\tau)\,\Delta\tau = \frac{w(\lambda, \tau)\,\Delta\tau}{1 - F(\lambda, \tau)}, \qquad (5\text{-}122)$$

d'où l'on déduit

$$h(\tau) = \frac{w(\lambda, \tau)}{1 - F(\lambda, \tau)}. \qquad (5\text{-}123)$$

Pour la distribution exponentielle, on a:

$$h(\tau) = \frac{\lambda e^{-\lambda\tau}}{e^{-\lambda\tau}} = \lambda, \qquad (5\text{-}124)$$

ce qui signifie que la fonction de risque est une constante égale à l'intensité des pannes.

L'expérience d'utilisation de l'appareillage électronique enseigne que la courbe représentative de la fonction $h(\tau)$ a l'allure de la figure 5-8. La portion initiale de cette courbe accuse une intensité des pannes accrue due à ce que certains éléments possèdent des défauts cachés qui se manifestent pendant la période initiale de fonctionnement. A mesure que τ devient élevé, l'intensité des pannes croît également,

ce qui est dû au « vieillissement » des éléments qui s'exprime par la dégradation de la qualité des éléments après un emploi prolongé. L'intensité des pannes constante et, partant, la distribution exponentielle n'ont lieu que dans la portion moyenne de la courbe $h(\tau)$.

On arrive cependant à supprimer l'intensité des pannes accrue dans la période initiale en soumettant l'appareillage à un entraînement préliminaire avant la mise en service. D'autre part, la croissance ininterrompue des exigences envers la qualité de l'appareillage, résultant du progrès technique, fait que la qualité de l'appareillage cesse de satisfaire, à la longue, aux exigences nouvelles plus sévères. Il se produit le vieillissement moral de l'appareillage, en général, avant que le vieillissement physique des éléments se fasse sentir. Cela permet d'admettre, dans bien des cas, que la fonction $h(\tau)$ est constante et égale à λ pendant toute la durée de service de l'appareillage, c.-à-d. d'apprécier la fiabilité en se basant sur la loi de distribution exponentielle.

e) Chaînes markoviennes

Soit $Z = \{z_1, \ldots, z_L\}$ l'espace des épreuves ou l'espace des états d'un certain système, qui reste le même pour toute étape du processus stochastique. Il est alors commode de désigner les états du système simplement par des indices affectant z en entendant par j l'état $z_j \in Z$. L'espace des états se représentera dans ce cas par l'ensemble des indices $J = \{1, \ldots, L\}$.

Désignons par $\pi_n = (\pi_n^{(1)}, \ldots, \pi_n^{(L)})$ la distribution des probabilités sur l'ensemble J pour la n-ième étape. A ce moment $\pi_n^{(j)}$ définit la probabilité pour qu'à la n-ième étape le système soit dans l'état j. On a déjà indiqué que le processus stochastique est une chaîne markovienne homogène si les probabilités de transition du système p_{ij} de l'état i à l'état j ne dépendent que de l'état i à l'étape précédente et sont les mêmes pour toute étape.

Il y a deux méthodes de représentation des probabilités de transition. La première consiste à mettre les probabilités de transition sous forme d'un tableau appelé matrice de transition et noté **P**. Pour $L = 3$, cette matrice est de la forme

$$\mathbf{P} = \begin{Vmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{Vmatrix}. \qquad (5\text{-}125)$$

La condition (5-108) à laquelle doivent nécessairement satisfaire les probabilités de transition signifie que la somme des éléments dans chaque ligne de la matrice des transitions doit nécessairement être égale à l'unité.

La seconde méthode de représentation des probabilités de transition consiste à construire le diagramme de transition, dont on voit sur la figure 5-9, *a* un exemple pour un système tristable. Le diagramme des transitions représente un graphe dont les sommets correspondent aux états du système et les arcs orientés indiquent les transitions possibles d'un état à un autre. Les probabilités des transitions s'expriment par des nombres associés à chaque arc. Conformément à la

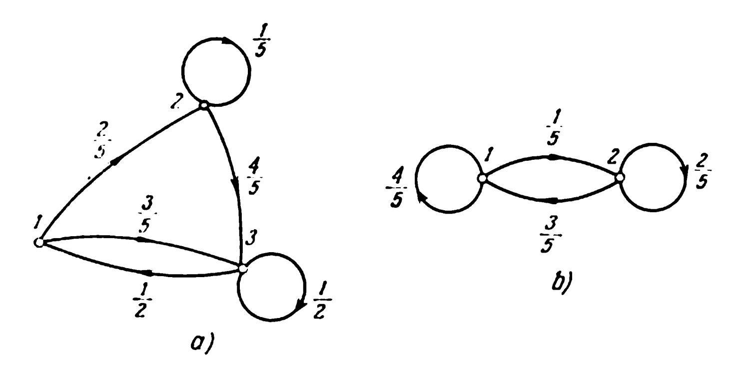


Fig. 5-9. Diagrammes des transitions pour une chaîne markovienne

condition (5-108), la somme des probabilités pour les arcs incidents vers l'extérieur à tout sommet du graphe doit être égale à l'unité.

Nombre de processus dans diverses branches de la science et de la technique peuvent être réduits aux chaînes markoviennes. En sociologie, les chaînes markoviennes facilitent l'étude des problèmes de variation de la structure sociale ou professionnelle de la population, des problèmes de migration de la population, etc. En biologie on étudie à l'aide des chaînes markoviennes le caractère d'évolution des diverses espèces d'animaux et de plantes. En physique, les chaînes markoviennes servent à l'étude de la diffusion de gaz. En technique, on se sert des chaînes markoviennes pour décrire certains processus de transmission de messages, différents processus de production, certains processus de contrôle de l'état d'utilisation et de recherche de défauts dans des dispositifs techniques complexes, et ainsi de suite.

Exemple 5-9. L'usine produit des téléviseurs d'une marque déterminée. Suivant que le type de téléviseur en question jouit ou non de la demande auprès des consommateurs, l'usine peut se trouver, à la fin de l'année, dans l'un des deux états: 1, la demande existe; 2, il n'y a pas de demande. La demande varie avec le temps, de sorte qu'on a la probabilité de 4/5 pour que vers la fin de l'année l'usine reste dans l'état 1. D'autre part, si l'usine s'est trouvée dans l'état 2, on prend des mesures en vue de modifier et de perfectionner le modèle, de sorte que vers la fin de l'année suivante l'usine passera à l'état 1 avec la probabilité de 3/5.

Dans l'exemple proposé, le développement de la production est représenté par une chaîne markovienne avec la matrice des transitions

$$\mathbf{P}=\begin{Vmatrix} \frac{4}{5} & \frac{1}{5} \\ \frac{3}{5} & \frac{2}{5} \end{Vmatrix}.$$

Le diagramme des transitions est donné sur la figure 5-9,*b*.

Quand on étudie des chaînes markoviennes, on s'attache avant tout à savoir la variation de la distribution des probabilités des états π_n lorsque le système passe d'un état à un autre. Désignons par T l'événement consistant en ce que le système se trouve à la $(n+1)$-ième étape dans l'état j.

Dans les notations adoptées, la probabilité de cet événement sera $P(T)=\pi_{n+1}^{(j)}$. Désignons par S_i l'événement consistant en ce que le système prenne l'état i à la n-ième étape, de sorte que $P(S_i)=\pi_n^{(i)}$. La grandeur p_{ij} est alors la probabilité pour que le système se trouve à la $(n+1)$-ième étape dans l'état j si à la n-ième étape il a été dans l'état i, c.-à-d. $p_{ij}=P(T\mid S_i)$. D'après la formule de la probabilité totale (5-42), on trouve:

$$\pi_{n+1}^{(j)}=\sum_{i=1}^{L}\pi_n^{(i)}p_{ij},\qquad j=1,\ \ldots,\ L. \tag{5-126}$$

La formule (5-126) permet de déterminer successivement, étape après étape, la variation de la distribution des probabilités des états du système si la distribution initiale des probabilités est connue.

Exemple 5-10. Supposons que l'usine (exemple 5-9) se trouve à l'instant initial dans l'état 1, c.-à-d. la distribution initiale des probabilités a la forme $\pi_0=(1,\ 0)$. Utilisant la formule (5-126) et la matrice des transitions, on a pour la première étape:

$$\pi_1^{(1)}=\pi_0^{(1)}p_{11}+\pi_0^{(2)}p_{21}=1\cdot\frac{4}{5}+0\cdot\frac{3}{5}=0{,}8\ ;$$

$$\pi_1^{(2)}=\pi_0^{(1)}p_{12}+\pi_0^{(2)}p_{22}=1\cdot\frac{1}{5}+0\cdot\frac{2}{5}=0{,}2.$$

Pour la deuxième étape

$$\pi_2^{(1)}=0{,}76,\qquad \pi_2^{(2)}=0{,}24.$$

Poursuivant ces calculs, on obtient la variation successive de la distribution des probabilités définie par le tableau 5-1 pour l'état initial (1, 0). On voit de ce tableau qu'à mesure que n croît, $\pi_n^{(1)}$ tend vers 0,75, et $\pi_n^{(2)}$, vers 0,25.

Si l'état initial du système est l'état 2, de sorte que $\pi_0=(0,\ 1)$, la distribution des probabilités aux étapes successives aura la forme donnée par le tableau 5-1. Dans ce cas, avec l'accroissement de n, les grandeurs $\pi_n^{(1)}$ et $\pi_n^{(2)}$ tendent vers les mêmes valeurs 0,75 et 0,25 qu'avec la distribution initiale $\pi_0=(1,\ 0)$. On voit donc que dans la chaîne markovienne donnée les probabilités des transitions après un grand nombre d'étapes parcourues deviennent indépendantes de l'état initial du système.

Tableau 5-1

Variation de la distribution des probabilités des états du processus markovien

n	Etat initial $\pi_0 = (1, 0)$						
	0	1	2	3	4	5	...
$\pi_n^{(1)}$	1	0,8	0,76	0,752	0,7504	0,75008	...
$\pi_n^{(2)}$	1	0,2	0,24	0,248	0,2498	0,24992	...
n	Etat initial $\pi_0 = (0, 1)$						
	0	1	2	3	4	5	...
$\pi_n^{(1)}$	0	0,6	0,72	0,744	0,7488	0,74976	...
$\pi_n^{(2)}$	1	0,4	0,28	0,256	0,2512	0,25024	...

S'il existe, pour une chaîne markovienne, une distribution limite des probabilités correspondant à $n \to \infty$ et indépendante de l'état initial du système, cette distribution définit le régime limite ou *établi* du système. On dit alors que le système est *statiquement stable*, et le processus markovien se produisant dans ce système est dit *ergodique*.

Désignons par $\pi = (\pi^{(1)}, \ldots, \pi^{(L)})$ la distribution établie des probabilités dans la chaîne markovienne ergodique. Les composantes $\pi^{(j)}$ de cette distribution peuvent être déduites des équations (5-126), qui deviennent pour $n \to \infty$:

$$\pi^{(j)} = \sum_{i=1}^{L} \pi^{(i)} p_{ij}, \qquad j = 1, \ldots, L. \tag{5-127}$$

Or, les L équations (5-127) ne sont pas toutes linéairement indépendantes, car les probabilités $\pi^{(j)}$ sont liées entre elles par la relation

$$\sum_{j=1}^{L} \pi^{(j)} = 1. \tag{5-128}$$

Aussi, pour déterminer L composantes inconnues $\pi^{(j)}$ de la distribution des probabilités π, suffit-il de prendre $L - 1$ équations quelconques (5-127) et de les résoudre conjointement avec l'équation (5-128).

Exemple 5-11. Pour la matrice des transitions de l'exemple 5-9, la première des équations (5-127) a la forme:

$$\pi^{(1)}=\pi^{(1)}p_{11}+\pi^{(2)}p_{21}=\frac{4}{5}\pi^{(1)}+\frac{3}{5}\pi^{(2)}.$$

L'équation (5-128) donne:

$$\pi^{(1)}+\pi^{(2)}=1.$$

Résolvant ces deux équations ensemble, on trouve:

$$\pi^{(1)}=0{,}75\,;\ \pi^{(2)}=0{,}25.$$

Le processus de passage d'une chaîne markovienne ergodique de l'état initial au régime établi porte le nom de *processus transitoire*. Le processus transitoire se décrit par la suite des distributions des probabilités π_n pour $n = 1, 2, \ldots$; la distribution initiale des probabilités π_0 étant connue, on construit ce processus par application successive de la formule (5-126), comme on l'a fait dans l'exemple 5-8. On peut aussi faire autrement.

Introduisons les quantités $p_{ij}^{(l)}$ définissant la probabilité de passage du système de l'état i à l'état j au cours de l étapes:

$$p_{ij}^{(l)}=p\,(x_{n+l}=j\,|\,x_n=i),\quad n=0,\,1\,\ldots \tag{5-129}$$

Connaissant les probabilités $p_{ij}^{(l)}$, on arrive à rechercher la distribution des probabilités π_l d'après la distribution π_0 par application d'une formule analogue à (5-126):

$$\pi_l^{(j)}=\sum_{i=1}^{L}\pi_0^{(i)}p_{ij}^{(l)},\quad j=1,\ldots,L. \tag{5-130}$$

Pour déterminer la liaison des probabilités de transition $p_{ij}^{(l)}$ avec les probabilités p_{ij}, posons $l = l_1 + l_2$ et considérons le passage du système de l'état π_0 à l'état π_l en deux stades: passage de π_0 à π_{l_1}, puis passage de π_{l_1} à $\pi_l = \pi_{l_1+l_2}$. On a alors

$$\pi_l^{(j)}=\sum_{k=1}^{L}\pi_{l_1}^{(k)}p_{kj}^{(l_2)}, \tag{5-131}$$

où

$$\pi_{l_1}^{(k)}=\sum_{i=1}^{L}\pi_0^{(i)}p_{ik}^{(l_1)}. \tag{5-132}$$

Substituant la valeur de $\pi_{l_1}^{(k)}$ dans (5-131), on obtient:

$$\pi_l^{(j)}=\sum_{k=1}^{L}p_{kj}^{(l_2)}\sum_{i=1}^{L}\pi_0^{(i)}p_{ik}^{(l_1)}=\sum_{i=1}^{L}\pi_0^{(i)}\sum_{k=1}^{L}p_{ik}^{(l_1)}p_{kj}^{(l_2)}. \tag{5-133}$$

Comparons (5-133) avec (5-130); il vient:

$$p_{ij}^{(l_1+l_2)}=\sum_{k=1}^{L}p_{ik}^{(l_1)}p_{kj}^{(l_2)}. \tag{5-134}$$

Pour le calcul successif des probabilités de transition $p_{ij}^{(1)}$, $p_{ij}^{(2)}$,. . . il est commode de poser dans (5-134) $l_1 = l - 1$, $l_2 = 1$. Il vient alors:

$$p_{ij}^{(l)} = \sum_{k=1}^{L} p_{ik}^{(l-1)} p_{ki}. \qquad (5\text{-}135)$$

Posant successivement $l = 1, 2, \ldots$ et se rappelant que par définition des probabilités de transition p_{ij}

$$p_{ik}^{0} = \begin{cases} 1, & i = k; \\ 0, & i \neq k, \end{cases} \qquad (5\text{-}136)$$

on trouve:

$$p_{ij}^{(1)} = p_{ij}, \qquad p_{ij}^{(2)} = \sum_{k=1}^{L} p_{ik} p_{kj}, \ldots \qquad (5\text{-}137)$$

L'expression (5-135) peut être généralisée au cas d'une chaîne markovienne non homogène, dans laquelle les probabilités de transition dépendent de l'indice de l'étape. Dans ce cas on entend par probabilité $p_{ij}(n, n_1)$ pour $n < n_1$ la probabilité pour qu'à l'étape n_1 le système se trouve dans l'état j si à l'étape n il était dans l'état i:

$$p_{ij}(n, n_1) = p(x_{n_1} = j \mid x_n = i), \qquad n < n_1. \qquad (5\text{-}138)$$

Par analogie avec (5-130), on a alors:

$$\pi_{n_1}^{(j)} = \sum_{i=1}^{L} \pi_n^{(i)} p_{ij}(n, n_1), \qquad n < n_1. \qquad (5\text{-}139)$$

D'autre part, considérant un certain n' vérifiant la condition $n < n' < n_1$, on arrive à mettre $\pi_{n_1}^{(j)}$ sous la forme

$$\pi_{n_1}^{(j)} = \sum_{k=1}^{L} \pi_{n'}^{(k)} p_{kj}(n', n_1) = \sum_{k=1}^{L} p_{kj}(n', n_1) \sum_{i=1}^{L} \pi_n^{(i)} p_{ik}(n, n') =$$
$$= \sum_{i=1}^{L} \pi_n^{(i)} \sum_{k=1}^{L} p_{ik}(n, n') p_{kj}(n', n_1). \qquad (5\text{-}140)$$

Par comparaison de (5-139) avec (5-140), on constate que

$$p_{ij}(n, n_1) = \sum_{k=1}^{L} p_{ik}(n, n') p_{kj}(n', n) \qquad (5\text{-}141)$$

pour tout $n < n' < n_1$. La relation (5-141) porte le nom d'*équation de Chapman-Kolmogorov.*

5-7. ÉLÉMENTS DE STATISTIQUE MATHÉMATIQUE

a) Objet de la statistique mathématique

L'évolution de la science et de la technique exige la pénétration de plus en plus profonde dans l'essence des phénomènes de la nature. Or ces phénomènes eux-mêmes se présentent à nos yeux sous la forme d'une multitude de faits et observations très variés, qui résultent à leur tour de l'action de très nombreux facteurs; une partie de ces facteurs sont réellement la cause du phénomène en question, tandis que d'autres, auxiliaires et insignifiants, ne font souvent que voiler le contenu du phénomène. De grandes connaissances théoriques et pratiques sont nécessaires pour éliminer toute l'information secondaire et mettre en valeur les éléments fondamentaux et essentiels dispersés dans les observations.

Utilisant les méthodes de la statistique mathématique, on arrive à mettre la multitude des résultats de l'observation sous une forme compacte, se prêtant à l'étude. Ces méthodes permettent de tirer l'information essentielle de l'ensemble des observations en la résumant dans quelques paramètres généralisés. S'il se trouve que les données collectées ne suffisent pas pour aller au fond du phénomène et qu'une expérience supplémentaire s'impose, il y a des méthodes statistiques par lesquelles on arrive à effectuer cette expérience de façon optimale, de sorte que le travail du chercheur soit facilité au maximum tant à la réalisation de l'expérience qu'au dépouillement de ses résultats.

On conclut de ce qui précède que la statistique mathématique est la science des méthodes de dépouillement d'une grande quantité de données expérimentales visant à en dégager des conclusions correctes.

Il est impossible d'examiner dans ce chapitre tous les problèmes relevant des méthodes de la statistique mathématique; nombre de manuels et de monographies [35 à 38] y sont consacrés. Aussi se bornera-t-on à considérer les méthodes les plus employées de résolution de problèmes statistiques. Quelques-unes de ces méthodes seront développées au chapitre 9 sous l'optique de la théorie de la prise de décisions.

b) Notion d'échantillon aléatoire

Soit x une variable aléatoire unidimensionnelle dont la fonction de répartition est $F(x)$. Considérons une variable aléatoire à n dimensions

$$(x^{(1)}, \ldots, x^{(n)}) \tag{5-142}$$

telle que ses différentes composantes soient des variables aléatoires indépendantes ayant les mêmes fonctions de répartition $F(x)$. Une telle variable multidimensionnelle aura pour fonction de ré-

partition le produit de celles de ses composantes

$$\prod_{i=1}^{n} F(x^{(i)}). \qquad (5\text{-}143)$$

Il est commode de considérer les variables aléatoires $x^{(i)}$ comme des épreuves, ou résultats d'une certaine expérience. Dans ce cas la variable aléatoire multidimensionnelle (5-142) peut être assimilée ou bien au résultat de n expériences indépendantes effectuées successivement sur le même matériel ou bien au résultat de n expériences effectuées simultanément sur n matériels de même type.

Etant donné qu'on peut toujours — du moins en théorie — faire un nombre infini d'expériences, il s'agit, au fond, d'une collection infinie de variables aléatoires $x^{(i)}$ dont la fonction de répartition est $F(x)$. Une telle collection infinie porte le nom de *population infinie à fonction de répartition* $F(x)$.

Chaque épreuve concrète, c.-à-d. chaque $x^{(i)}$ faisant partie de la collection finie (5-142), peut être considérée comme le choix d'un nombre sur une population infinie. La collection (5-142) tout entière représente la suite de n choix de cette espèce et s'appelle *échantillon aléatoire*.

c) Théorèmes limites de la théorie des probabilités

En statistique mathématique, on accorde un grand intérêt à l'étude de la variation des propriétés de l'échantillon aléatoire avec l'augmentation de sa taille et, plus particulièrement, aux relations limites qu'on obtient quand $n \to \infty$. Dans cette étude, le rôle primordial revient à l'inégalité de Tchébychev, à la loi des grands nombres qui découle de cette dernière et au théorème limite central.

Si x est une variable aléatoire d'espérance mathématique ν et de variance σ^2, elle vérifie l'inégalité suivante dite *inégalité de Tchébychev*:

$$P(|x-\nu| \geqslant \lambda\sigma) \leqslant \frac{1}{\lambda^2}. \qquad (5\text{-}144)$$

Pour démontrer cette inégalité, partageons l'axe réel en trois intervalles:

$$I - (-\infty,\ \nu - \lambda\sigma];\ I' = (\nu - \lambda\sigma,\ \nu + \lambda\sigma);$$
$$I'' = [\nu + \lambda\sigma,\ +\infty).$$

La variance de x peut s'écrire comme

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\nu)^2 w(x)\, dx = \int_I (x-\nu)^2 w(x)\, dx + \\ + \int_{I'} (x-\nu)^2 w(x)\, dx + \int_{I''} (x-\nu)^2 w(x)\, dx.$$

Supprimant dans l'expression de σ^2 le terme répondant à l'intervalle I', on a:

$$\sigma^2 \leqslant \int_{I} (x-\nu)^2 w(x)\,dx + \int_{I''} (x-\nu)^2 w(x)\,dx.$$

Comme dans l'intervalle I on a $x - \nu \leqslant -\lambda\sigma$, et dans l'intervalle I'' $x - \nu \geqslant \lambda\sigma$, l'inégalité ne sera que plus forte si l'on remplace dans les deux intégrales $(x-\nu)^2$ par $\lambda^2\sigma^2$. De cette façon,

$$\sigma^2 \leqslant \lambda^2\sigma^2 [P(x \leqslant \nu - \lambda\sigma) + P(x \geqslant \nu + \lambda\sigma)] =$$
$$= \lambda^2\sigma^2 P(|x-\nu| \geqslant \lambda\sigma),$$

ce qui équivaut à (5-144).

L'inégalité (5-144) définit la probabilité pour que la grandeur $x - \nu$ dépasse les limites de l'intervalle $(-\lambda\sigma, +\lambda\sigma)$. Or, on peut aussi déterminer la probabilité pour que $x - \nu$ reste dans les limites de l'invervalle indiqué; cette probabilité est

$$P(|x-\nu| < \lambda\sigma) = 1 - P(|x-\nu| \geqslant \lambda\sigma),$$

ou, compte tenu de (5-144),

$$P(|x-\nu| < \lambda\sigma) \geqslant 1 - \frac{1}{\lambda^2}. \tag{5-145}$$

Désignons $\lambda\sigma$ par ε. L'inégalité devient

$$P(|x-\nu| < \varepsilon) \geqslant 1 - \frac{\sigma^2}{\varepsilon^2}. \tag{5-146}$$

Revenons à l'échantillon aléatoire $(x_1, \ldots, x_n)$ de taille n. Désignons par $\bar{x}$ la moyenne arithmétique de cet échantillon qui est égale à

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{5-147}$$

Afin de simplifier l'écriture, nous adopterons dans la suite la notation S ou S_n pour désigner la sommation sur toutes les données d'échantillonnage. La relation (5-147) s'écrira alors

$$\bar{x} = \frac{1}{n} S(x) = \frac{S_n}{n}. \tag{5-148}$$

Conformément à (5-85) et à (5-100), la variance de $\bar{x}$ est

$$D(\bar{x}) = D\left[\frac{S(x)}{n}\right] = \frac{1}{n^2} S[D(x)] = \frac{\sigma^2}{n}. \tag{5-149}$$

Remplaçant dans (5-146) x par $\bar{x}$ et, partant, σ^2 par σ^2/n, on obtient:

$$P\left\{\left|\frac{S_n}{n} - \nu\right| < \varepsilon\right\} \geqslant 1 - \frac{\sigma^2}{n\varepsilon^2}, \tag{5-150}$$

où ν est l'espérance mathématique de $\bar{x}$, qui est une valeur moyenne pour la population infinie soumise à l'échantillonnage.

On voit de (5-150) que pour $n \to \infty$ et pour tout $\varepsilon > 0$

$$P\left\{\left|\frac{S_n}{n} - \nu\right| < \varepsilon\right\} \to 1; \tag{5-151}$$

c'est la *loi des grands nombres*. Il en découle que, pour des n grands, on a la probabilité aussi proche de l'unité que l'on veut pour que la moyenne arithmétique et l'espérance mathématique $\nu = M(x)$ d'une variable x se confondent.

Citons sans démonstration une des plus fondamentales relations limites de la théorie des probabilités: c'est le *théorème limite central.* Soit $(x_1, \ldots, x_n)$ une suite de variables aléatoires qui ont, en général, les fonctions de répartition $F_1(x), \ldots, F_n(x)$, les moyennes $\nu_1, \ldots, \nu_n$ et les variances $\sigma_1^2, \ldots, \sigma_n^2$ différentes. Adoptons les notations

$$\bar{x} = \frac{1}{n} S(x); \quad \nu = \frac{1}{n} S(\nu); \quad s^2 = \frac{1}{n} S(\sigma^2). \tag{5-152}$$

Le théorème limite central stipule que si les variables aléatoires $x_1, \ldots, x_n$ sont indépendantes et ont des variances finies de même ordre, leur moyenne arithmétique x tend (quand le nombre de termes est grand) vers la distribution normale caractérisée par la moyenne ν et la variance s^2. Remarquant que la variable

$$t = \frac{\bar{x} - \nu}{s/\sqrt{n}} \tag{5-153}$$

aura alors la moyenne égale à zéro et la variance égale à l'unité, on peut affirmer en se basant sur le théorème limite central que pour $n \to \infty$ l'on aura

$$P\left(\frac{\bar{x} - \nu}{s/\sqrt{n}} < y\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{1}{2}u^2} du. \tag{5-154}$$

Les problèmes dans lesquels la variable aléatoire étudiée représente la somme d'un grand nombre de variables aléatoires indépendantes sont très fréquents. C'est ainsi que l'erreur d'un instrument compliqué est le résultat de l'addition de différentes sollicitations extérieures et d'erreurs fournies par les éléments isolés de l'instrument en question. Dans ce cas les erreurs dont l'influence est notable sont faciles à éliminer.

Si une coupure se produit dans un circuit électrique, elle est facile à localiser et à éliminer. Par contre, il est difficile de mettre au clair les causes des erreurs de faible grandeur. Donc, considérant les composantes isolées de l'erreur totale comme uniformément petites, on peut affirmer, sur la base du théorème limite central, que la distribution de l'erreur totale est proche de la distribution normale.

d) Problèmes de la statistique mathématique

Le problème fondamental de la statistique mathématique est la recherche des paramètres inconnus de la distribution d'une population infinie d'après un échantillon fini connu. On peut chercher alors tant la fonction de répartition $F(x)$ elle-même que des paramètres isolés de celle-ci tels que la moyenne, la variance, la plus grande ou la plus petite valeur de la variable aléatoire, etc.

Les éléments de l'échantillon fini étant des variables aléatoires, la valeur du paramètre recherchée d'après cet échantillon sera elle aussi une variable aléatoire. Plus particulièrement, disposant de plusieurs échantillons de même taille faits sur une seule et même population infinie, on obtient autant de valeurs du paramètre en question qu'il y a d'échantillons. Aussi l'échantillon fini ne permet-il pas de savoir avec exactitude la valeur du paramètre, mais seulement d'en faire une estimation plus ou moins précise. Les valeurs numériques des différents paramètres tirées d'un échantillon fini portent le nom d'*estimateurs* des paramètres de la population infinie.

Dans le cas général, on peut tirer des estimateurs différents d'un seul et même échantillon. Soit, par exemple, un échantillon $\{x_1, x_2, \ldots, x_n\}$ tel que $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$; on demande de savoir l'estimateur de la moyenne. Conformément aux formules (3-24), (3-25) et (3-27), la moyenne admet plusieurs estimateurs. Ce sont:

la moyenne de toutes les observations, ou *médiane*:

$$\hat{\bar{x}} = \begin{cases} x_{\frac{n+1}{2}} & \text{pour } n \text{ impair;} \\ \frac{1}{2}\left(x_{\frac{n}{2}-1} + x_{\frac{n}{2}+1}\right) & \text{pour } n \text{ pair;} \end{cases} \tag{5-155}$$

la demi-somme de la plus petite et de la plus grande valeur

$$\hat{\bar{x}} = \frac{1}{2}(x_1 + x_n); \tag{5-156}$$

la moyenne arithmétique de toutes les observations

$$\hat{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{5-157}$$

Lequel de ces estimateurs est-il à préférer? La réponse à cette question est donnée en introduisant certains critères qui doivent être remplis par l'estimateur. Considérons-en les plus importants.

1. L'estimateur doit être *sans biais*, c.-à-d. exempt d'erreur systématique susceptible de faire surestimer ou sous-estimer la valeur du paramètre dans tous les échantillons. Cela signifie que l'espérance mathématique de l'estimateur doit être égale à la valeur réelle du paramètre. Désignant la valeur réelle du paramètre par α et son

estimateur par $\hat{\alpha}$, on exprime l'absence de biais en écrivant

$$M(\hat{\alpha}) = \alpha. \tag{5-158}$$

2. L'estimateur doit être *consistant*, c.-à-d. qu'il doit tendre vers la valeur du paramètre α à mesure que la taille de l'échantillon croît. L'estimateur $\hat{\alpha}$ étant une variable aléatoire, cette tendance aura nécessairement un caractère probabiliste. Si l'on désigne par $\hat{\alpha}_n$ l'estimateur de α obtenu d'après un échantillon de taille n, l'estimateur consistant doit vérifier la relation

$$P[|\hat{\alpha}_n - \alpha| < \varepsilon] \to 1 \tag{5-159}$$

pour $n \to \infty$ et pour tout $\varepsilon > 0$.

3. L'estimateur doit être *efficace*, c.-à-d. qu'il doit être le plus proche du paramètre à estimer que n'importe quel autre estimateur consistant et sans biais. En d'autres mots, il doit minimiser le nombre d'écarts excessifs à l'utilisation des différents échantillons. Mathématiquement, cela équivaut à l'exigence du minimum de la variance de l'estimateur:

$$D(\hat{\alpha} - \alpha) = \min. \tag{5-160}$$

e) Estimateurs sans biais de la moyenne et de la variance

Soit $(x_1, \ldots, x_n)$ un échantillon d'une population de moyenne ν et de variance σ^2. Désignons par $\bar{x}$ l'estimateur de la moyenne, et par s^2, celui de la variance.

Pour l'estimateur de ν on prend généralement la moyenne arithmétique de l'échantillon, ou *moyenne d'échantillon*,

$$x = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} S(x). \tag{5-161}$$

En vertu de la loi des grands nombres, un tel estimateur est consistant, c.-à-d. il tend (converge en probabilité) vers ν quand $n \to \infty$.

Puisqu'elle représente la somme des variables aléatoires, la moyenne d'échantillon $\bar{x}$ est elle-même une variable aléatoire, caractérisée par une loi de distribution, une moyenne et une variance.

La moyenne, ou l'espérance mathématique de la moyenne d'échantillon, est égale à

$$M(\bar{x}) = \frac{1}{n} M[S(x)] = \frac{1}{n} S[M(x)] = \nu. \tag{5-162}$$

Nous voyons que l'espérance mathématique de la moyenne d'échantillon coïncide avec la valeur de ν, ce qui témoigne de l'absence de biais de l'estimateur (5-161). Conformément à (5-149), la variance de la moyenne d'échantillon est σ^2/n.

Si l'échantillon est fait sur une population à distribution normale, alors $\bar{x}$, en tant que somme de variables aléatoires à distribution normale, aura aussi la distribution normale qui, compte tenu de (5-162) et de (5-149), aura la forme

$$w(\bar{x}) = N\left(\nu, \frac{\sigma^2}{n}\right). \tag{5-163}$$

Dans bien des cas, il est commode de considérer non pas $\bar{x}$ mais la variable

$$t = \frac{\bar{x}-\nu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\,(\bar{x}-\nu)}{\sigma} \tag{5-164}$$

ayant la moyenne égale à zéro et la variance égale à l'unité; cette variable a la densité de probabilité $N(0, 1)$.

Si l'échantillon est fait sur une population dont la distribution n'est pas normale, la loi de distribution de la moyenne d'échantillon ne sera normale non plus; or, en vertu du théorème limite central, on peut admettre que pour les échantillons de grande taille cette loi est proche de la loi $N(\nu, \sigma^2/n)$.

Pour trouver l'estimateur sans biais de la variance s^2, cherchons l'espérance mathématique de la variable $S(x - \bar{x})^2$. Mettant cette variable sous la forme

$$S(x-\bar{x})^2 = S[(x-\nu)-(\bar{x}-\nu)]^2 =$$
$$= S(x-\nu)^2 - 2(\bar{x}-\nu)\,S(x-\nu) + n(\bar{x}-\nu)^2$$

et prenant l'espérance mathématique de chaque terme, on obtient:

$$M[S(x-\bar{x})^2] = (n-1)\,\sigma^2 \tag{5-165}$$

ou

$$M\left[\frac{S(x-\bar{x})^2}{n-1}\right] = \sigma^2. \tag{5-166}$$

Il découle de (5-166) que l'estimateur sans biais de la variance est un estimateur de la forme

$$s^2 = \frac{1}{n-1}\,S(x-\bar{x})^2. \tag{5-167}$$

Avant de définir la loi de distribution de l'estimateur de la variance, considérons une loi de distribution très importante. Si y_i sont des variables aléatoires indépendantes obéissant à la distribution normale $N(0, 1)$, la variable aléatoire

$$\chi^2 = \sum_{i=1}^{n} y_i^2 \tag{5-168}$$

a la distribution χ^2 avec n degrés de liberté, le nombre n étant déterminé par le nombre de variables aléatoires indépendantes formant

la somme (5-168). Il existe pour la distribution χ^2 des tables très détaillées, dans lesquelles on fournit pour différents n les valeurs de la probabilité $p\ (\chi^2 - \chi_q^2)$, où χ_q^2 est un nombre positif quelconque.

Par exemple, le tableau 5-2 donne les valeurs de χ_q^2 vérifiant la condition $p\ (\chi^2 - \chi_q^2) = 0{,}95$ pour des n de 1 à 30. On trouve dans [36] des tables plus détaillées de la distribution χ_q^2.

Tableau 5-2

Valeurs de la borne supérieure de 95 % de χ_q^2 pour la distribution χ^2 en fonction du nombre n de degrés de liberté

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
χ_q^2	0,004	0,103	0,352	0,71	1,14	1,63	2,17	2,73	3,32	3,94	4,6	5,2	5,9	6,6	7,3
n	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
χ_q^2	8,0	8,7	9,4	10,1	10,9	11,6	12,3	13,1	13,8	14,6	15,4	16,2	16,9	17,7	18,5

Il est facile de voir que si y_i est la variable $(x_i - x)/\sigma$ vérifiant les conditions imposées à y_i, la grandeur

$$\chi^2 = \sum_{i=1}^{n} y_i^2 = \frac{(n-1)\,s^2}{\sigma^2} \qquad (5\text{-}169)$$

définira la loi de distribution de s^2.

f) Estimation par le maximum de vraisemblance

Bon nombre de problèmes de statistique mathématique se réduisent à estimer un certain paramètre α d'une distribution dont on connaît la forme. Le principe du maximum de vraisemblance consiste à retenir en qualité d'estimateur du paramètre α la valeur qu'on juge la plus probable sur la base des données expérimentales.

Soit un échantillon $(x_1, \ldots, x_n)$ de la population à densité de probabilité $w\ (x, \alpha)$ exprimée en fonction du paramètre α. La densité multidimensionnelle de probabilité pour cet échantillon

$$L = \prod_{i=1}^{n} w\,(x_i, \alpha) \qquad (5\text{-}170)$$

porte le nom de *fonction de vraisemblance*. Conformément au principe du maximum de vraisemblance, on adopte comme estimateur de α

la valeur $\hat{\alpha}$ pour laquelle la fonction de vraisemblance L, ou, ce qui revient au même, son logarithme $\ln L$, admet un maximum. On détermine la valeur de $\hat{\alpha}$ à partir de l'équation de vraisemblance

$$\frac{\partial \ln L}{\partial \alpha}=\sum_{i=1}^{n} \frac{\partial \ln w(x_i, \alpha)}{\partial \alpha}=0 \tag{5-171}$$

qu'on réduit facilement à la forme

$$\sum_{i=1}^{n} \frac{1}{w(x_i, \alpha)} \frac{\partial w(x_i, \alpha)}{\partial \alpha}=0. \tag{5-172}$$

Dans le cas où l'échantillon est fait sur une population à distribution normale $N(\nu, \sigma^2)$ définie par (5-54), les équations de vraisemblance définissant les paramètres ν et σ ont la forme

$$\left.\begin{aligned} \sum_{i=1}^{n}(x_i-\nu)=0; \\ \sum_{i=1}^{n}\left[1-\frac{(x_i-\nu)^2}{\sigma^2}\right]=0. \end{aligned}\right\} \tag{5-173}$$

De la première équation on tire l'estimateur de ν

$$\hat{\nu}=\frac{1}{n}\sum_{i=1}^{n} x_i=\bar{x} \tag{5-174}$$

qui coïncide avec l'estimateur consistant sans biais (5-161) obtenu auparavant. La seconde équation de vraisemblance, conjointement avec (5-174), donne l'estimateur de σ^2

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{5-175}$$

qui a cependant un biais. Aussi emploie-t-on généralement, en pratique, l'estimateur sans biais (5-167).

Etudions sur un exemple la recherche de l'estimateur par maximum de vraisemblance pour une distribution exponentielle.

Exemple 5-12. Pour déterminer la durée de vie moyenne de l'appareillage produit par la fabrique, on a procédé à l'essai de n échantillons d'appareillage au cours de t heures. Lors de l'essai m échantillons ont fonctionné sans défaillance pendant toute la durée t, alors que $n-m$ échantillons sont tombés en panne aux instants $\tau_1, \ldots, \tau_{n-m}$. Quel est l'estimateur le plus vraisemblable du temps moyen de fonctionnement sans pannes?

Admettant que le temps pendant lequel il se produit une panne obéit à la loi exponentielle (5-120), de sorte que la grandeur $w(\lambda, \tau)=\lambda e^{-\lambda\tau}$ représente

la probabilité de la panne à l'instant τ et la grandeur $e^{-\lambda t}$, la probabilité de fonctionnement sans défaillance pendant la durée t, on obtient la fonction de vraisemblance sous la forme

$$L=(e^{-\lambda t})^m \prod_{i=1}^{n-m} w(\lambda, \tau_i). \qquad (5\text{-}176)$$

De l'équation (5-171) on tire alors:

$$-mt+\sum_{i=1}^{n-m}\left(\frac{1}{\lambda}-\tau_i\right)=0, \qquad (5\text{-}177)$$

d'où

$$\tau_{\text{moy}}=\frac{1}{\lambda}=\frac{mt+\sum\limits_{i=1}^{n-m}\tau_i}{n-m}. \qquad (5\text{-}178)$$

On montre de façon analogue que l'estimateur le plus vraisemblable de la probabilité d'un événement pour un processus d'essais indépendants et la distribution binomiale est égal à

$$\hat{p}=\frac{x}{n}, \qquad (5\text{-}179)$$

où n est le nombre d'essais et x le nombre de cas favorables.

Si le processus aléatoire obéit à la distribution de Poisson et que r événements soient contenus dans l'échantillon couvrant n intervalles, l'estimateur le plus vraisemblable du nombre moyen d'événements par intervalle est la grandeur

$$\hat{a}=\frac{r}{n}. \qquad (5\text{-}180)$$

g) Estimation des paramètres par la méthode des intervalles de confiance

Dans les paragraphes précédents on a étudié les méthodes fournissant l'estimation du paramètre sous la forme d'un nombre, c.-à-d. d'un point de l'axe réel. Une telle estimation est dite *ponctuelle*. Or, puisque l'échantillon est aléatoire, il arrive souvent, surtout quand la taille de l'échantillon est petite, que l'estimation obtenue s'écarte pour beaucoup de la vraie valeur du paramètre.

On obtient beaucoup plus d'information en indiquant l'intervalle dans lequel se situe, avec une probabilité proche de l'unité, la valeur du paramètre à estimer. Par exemple, si γ est une quantité proche de l'unité, l'intervalle (α_1, α_2) dans lequel se trouve, avec la probabilité γ, la valeur du paramètre inconnu α

$$P(\alpha_1<\alpha<\alpha_2)=\gamma \qquad (5\text{-}181)$$

porte le nom d'*intervalle de confiance* à $\gamma=100\%$. Pratiquement, on se contente souvent de rechercher l'intervalle de confiance à 95 % en posant $\gamma=0,95$. Après avoir trouvé l'estimateur ponctuel $\hat{\alpha}$ du pa-

ramètre et la densité de probabilité de cet estimateur, on tire les limites de l'intervalle de confiance de la relation

$$\int_{\alpha_1}^{\alpha_2} w(\hat{\alpha})\, d\hat{\alpha} = \gamma. \qquad (5\text{-}182)$$

Dans certains cas, au lieu de l'intervalle de confiance bilatéral (α_1, α_2), on ne demande de savoir que la limite inférieure α_1' ou supérieure α_2' du paramètre considéré α. Les intervalles définis par les relations

$$P(\alpha > \alpha_1') = \gamma \quad \text{et} \quad P(\alpha < \alpha_2') = \gamma \qquad (5\text{-}183)$$

sont dits *intervalles de confiance unilatéraux à* $\gamma = 100\%$. On établit les limites α_1' et α_2' de ces intervalles par les expressions

$$\int_{\alpha_1'}^{+\infty} w(\hat{\alpha})\, d\hat{\alpha} = \gamma \quad \text{et} \quad \int_{-\infty}^{+\alpha_2'} w(\hat{\alpha})\, d\hat{\alpha} = \gamma. \qquad (5\text{-}184)$$

Supposons que l'échantillon de taille n ait été pris sur une population à distribution normale $N(\nu, \sigma^2)$ avec ν inconnu et la variance σ^2 connue, de sorte que $t = \sqrt{n}(\bar{x} - \nu)/\sigma$ aura la distribution $N(0, 1)$. Introduisons encore la notation $\gamma = 1 - q$. Dans ce cas l'intervalle de confiance à $(1-q) = 100\%$ se définira par la relation

$$P\left(-t_q < \frac{\sqrt{n}\,(\bar{x} - \nu)}{\sigma} < +t_q\right) = 1 - q, \qquad (5\text{-}185)$$

ou

$$P\left(\bar{x} - t_q \frac{\sigma}{\sqrt{n}} < \nu < \bar{x} + t_q \frac{\sigma}{\sqrt{n}}\right) = 1 - q. \qquad (5\text{-}186)$$

Compte tenu de (5-58), on écrit

$$1 - q = \Phi(t_q) - \Phi(-t_q) = 2\Phi(t_q), \qquad (5\text{-}187)$$

en aboutissant à l'expression suivante de t_q:

$$\Phi(t_q) = \frac{1}{2}(1 - q). \qquad (5\text{-}188)$$

Pour les intervalles de confiance unilatéraux $(t_q', +\infty)$ et $(-\infty, t_q')$ la relation (5-187) s'écrira comme suit:

$$1 - q = \Phi(t_q') - \Phi(-\infty) = \frac{1}{2} + \Phi(t_q'), \qquad (5\text{-}189)$$

les valeurs limites des intervalles de confiance unilatéraux étant définies par

$$\Phi(t_q') = \frac{1}{2}(1 - 2q). \qquad (5\text{-}190)$$

On voit donc que les valeurs limites d'un intervalle de confiance unilatéral à $(1-q)=100$ % sont égales aux valeurs limites d'un intervalle de confiance bilatéral à $(1-q)=100$ %.

Posant $q = 0,05$ et consultant les tables de l'intégrale de probabilité, nous tirons des relations (5-188) et (5-190) les limites suivantes des intervalles de confiance bilatéraux et unilatéraux:

$$t_q = 1,96, \quad t'_q = 1,64. \tag{5-191}$$

Ainsi donc, connaissant la moyenne d'échantillon $\bar{x}$ calculée pour l'échantillon de taille n sur une population à distribution normale avec la variance σ^2 connue, on peut affirmer avec une probabilité de 0,95 que la moyenne ν de la population appartiendra à l'intervalle

$$\bar{x} - 1,96 \frac{\sigma}{\sqrt{n}} < \nu < \bar{x} + 1,96 \frac{\sigma}{\sqrt{n}}, \tag{5-192}$$

ou que

$$\nu > \bar{x} - 1,64 \frac{\sigma}{\sqrt{n}}, \tag{5-193}$$

ou que

$$\nu < \bar{x} + 1,64 \frac{\sigma}{\sqrt{n}}. \tag{5-194}$$

Les valeurs obtenues de t_q peuvent servir aussi à rechercher les intervalles de confiance pour la distribution binomiale $w(x, n, p)$, car, pour n suffisamment grands, la distribution binomiale tend vers la distribution normale caractérisée par la moyenne $\nu = np$ et la variance $\sigma^2 = npq$. Posant $t = (x - np)/\sqrt{npq}$, on aboutit à la conclusion selon laquelle, après x réalisations d'un événement S au cours de n essais indépendants, on a la probabilité 0,95 pour que la valeur moyenne np soit comprise entre les limites suivantes:

$$x - 1,96 \sqrt{npq} < np < x + 1,96 \sqrt{npq}, \tag{5-195}$$

ou pour que

$$np > x - 1,64 \sqrt{npq}, \tag{5-196}$$

ou pour que

$$np < x + 1,64 \sqrt{npq}, \tag{5-197}$$

la valeur de $\sqrt{npq}$ étant calculée d'après la valeur la plus vraisemblable de $p = \hat{p} = x/n$ comme

$$\sqrt{npq} = \sqrt{\frac{x(n-x)}{n}}. \tag{5-198}$$

Exemple 5-13. De la production de l'usine on a prélevé de façon aléatoire et essayé 100 articles dont 37 ont été signalés comme appartenant à la 2e qualité. Quel est le pourcentage de la 2e qualité dans tout le volume de production de l'usine?

Dans le cas donné $n = 100$, $x = 37$, $\sqrt{npq} = 4{,}84$. Alors $1{,}96\sqrt{npq} = 9{,}5$, de sorte que $x - 1{,}96\sqrt{npq} = 27{,}5$, $x + 1{,}96\sqrt{npq} = 46{,}5$.

De cette façon, il y a 5 chances sur 100 qu'on se trompe en faisant une des hypothèses suivantes qui disent que le nombre des articles de la 2e qualité

1) est compris entre 27,5 et 46,5 % de la totalité des articles;
2) n'est pas inférieur à 29 % de la totalité des articles;
3) n'est pas supérieur à 45 % de la totalité des articles.

On a très souvent recours aux intervalles de confiance pour estimer la grandeur ν pour une distribution normale si la variance σ^2 n'est pas connue. Dans ce cas on est amené à employer non pas σ^2 mais l'estimateur de la variance s^2. Or, la variable

$$t = \frac{\sqrt{n}(\bar{x} - \nu)}{s} \tag{5-199}$$

ne sera plus distribuée, à la différence de (5-164), suivant la loi normale. En réalité la variable t définie par (5-199) obéira à la loi de distribution de Student qui peut être définie de la façon suivante.

Soient u une variable aléatoire à distribution normale $N(0, 1)$ et v une autre variable aléatoire à distribution χ^2. Si u et v sont indépendantes, la variable aléatoire

$$t = \frac{u}{\sqrt{v/k}} = \frac{u\sqrt{k}}{\sqrt{v}} \tag{5-200}$$

définit la *distribution de Student* dans laquelle le nombre k porte le nom de nombre de degrés de liberté.

Définissant maintenant les variables

$$u = \frac{\sqrt{n}(\bar{x} - \nu)}{\sigma}; \quad v = \frac{(n-1)s^2}{\sigma^2}, \tag{5-201}$$

on s'assure qu'en vertu de (5-164) et de (5-169) ces variables vérifient les conditions définissant la distribution de Student et que la variable t déterminée d'après (5-200) coïncide avec la valeur fournie par (5-199) pour $k = n - 1$. De cette façon, la variable aléatoire t d'équation (5-199) sera distribuée suivant la loi de Student avec $k = n - 1$ degrés de liberté.

Désignons par t_{qk} les limites de l'intervalle de confiance bilatéral à $(1-q) = 100$ % pour la distribution de Student à k degrés de liberté. On trouve des tables détaillées (par exemple dans [36]) de t_{qk} pour différents q et pour k variant de 1 à 30. Lorsque $k > 30$, les valeurs de t_{qk} coïncident pratiquement avec les valeurs de t_q obtenues pour la distribution normale. On propose un exemple de table de cette espèce (tableau 5-3) fournissant les valeurs de t_{qk} pour $q = 0{,}05$ et $0{,}1$.

Tableau 5-3

Valeurs des limites de l'intervalle de confiance à $(1-q)=100\%$ en fonction du nombre de degrés de liberté k pour la distribution de Student

q	k										
	1	2	3	4	5	6	7	8	9	10	12
0,05	12,71	4,30	3,18	2,78	2,57	2,45	2,36	2,31	2,62	2,23	2,18
0,10	6,31	2,92	2,35	2,13	2,01	1,94	1,89	1,86	1,83	1,81	1,78

q	k									
	14	16	18	20	22	24	26	28	30	∞
0,05	2,15	2,12	2,10	2,09	2,07	2,06	2,06	2,05	2,04	1,96
0,10	1,76	1,75	1,73	1,72	1,72	1,71	1,71	1,70	1,70	1,64

On déduit les limites des intervalles de confiance unilatéraux t'_{qk} d'après les valeurs de t_{qk} de la relation

$$t'_{qk}=t_{2qk}. \qquad (5\text{-}202)$$

Exemple 5-14. Afin de réduire l'influence des parasites dans le canal de communication, chaque résultat de mesure d'un paramètre ν à bord d'un missile est transmis au sol trois fois. Les résultats des mesures sont $x_1 = 3{,}2$; $x_2 = 2{,}9$; $x_3 = 3{,}1$. On demande de définir l'intervalle de confiance bilatéral à 95 % pour le paramètre ν en admettant que les déformations subies par chaque valeur transmise sont réciproquement indépendantes et distribuées suivant la loi normale. D'après les données mesurées, on trouve:

$$n=3, \quad \bar{x}=\frac{1}{n}\,S\,(x)=3{,}067\,;$$

$$s^2=\frac{1}{n-1}\,S\,(x-\bar{x})^2=0{,}0234.$$

On lit dans le tableau 5-3 pour $q = 0{,}05$ et $k = n - 1 = 2$:

$$t_{qk} = 4{,}30, \text{ de sorte que } t_{qk}s/\sqrt{n} = 0{,}377.$$

Donc, on a la probabilité 0,95 pour que $2{,}69 < \nu < 3{,}44$.

h) Vérification des hypothèses statistiques. Notion de critère d'adéquation

Soit Z l'espace des épreuves à éléments $z \in Z$. Pour déterminer la probabilité $p\,(z)$ d'une épreuve z, on procède à n expériences. Si l'épreuve z a eu lieu n_z fois, on estime la probabilité de cette épreu-

ve d'après le rapport n_z/n. Or, en réalité, le rapport n_z/n nous donne non pas la probabilité $p(z)$ de l'épreuve z mais sa fréquence $q(z)$, qui peut être sensiblement différente de la probabilité, surtout quand le nombre d'expériences n'est pas grand. La question se pose: à quel point est-il légitime d'estimer la distribution des probabilités des épreuves $p(z)$ sur la base des fréquences $q(z)$ connues? Cependant, il n'est pas tout à fait correct de poser le problème de cette façon.

En effet, si l'on obtient à la suite de l'expérience quelques nombres exprimant les fréquences des différentes épreuves liées à l'expérience $q(z)$, il faut que la substitution à ces nombres d'autres nombres [notamment des probabilités $p(z)$] soit justifiée par quelque chose. Comment peut-on justifier le remplacement d'un nombre par un autre?

Supposons qu'en jetant une pièce de monnaie dix fois on a amené sept fois pile. La fréquence de pile dans l'expérience donnée est 0,7. Est-il légitime de modifier cette grandeur et de prendre un autre nombre déterminé comme probabilité de pile?

Si nous ne savons rien sur la pièce de monnaie, en particulier si nous ignorons qu'elle est symétrique ou non, nous n'avons aucune raison de modifier la valeur de la fréquence tirée de l'expérience. On peut cependant poser le problème autrement. On sait que si la pièce est symétrique, la probabilité de pile est 0,5. On veut voir si la pièce est bien symétrique; dans ce but on jette la pièce dix fois. On a sept fois pile. Cela veut-il dire que la pièce est symétrique?

Le problème ainsi formulé est le *problème de vérification de l'hypothèse*. De pareils problèmes se posent quand on a la possibilité de formuler une certaine hypothèse (par exemple que la pièce de monnaie est symétrique) portant sur le caractère de la distribution ou sur la valeur des paramètres de la distribution d'une variable aléatoire. Le but de l'expérience est de confirmer ou de rejeter l'hypothèse avancée. Les méthodes de la statistique mathématique permettent de donner la réponse à des questions comme celles-ci:

1) le nouveau modèle d'appareil a-t-il les qualités supérieures à celles des appareils existants?

2) la fiabilité du système ne sera-t-elle pas inférieure à celle requise?

3) la nouvelle méthode de cure sera-t-elle plus efficace que les méthodes existantes? etc.

Pour savoir si l'hypothèse avancée doit être acceptée ou rejetée, il faut élaborer un *critère d'adéquation* de l'hypothèse vérifiée aux résultats de l'expérience.

Supposons qu'on a des raisons de croire que la valeur d'un paramètre α est égale à α_0 et qu'on veut le vérifier expérimentalement. Après avoir effectué l'expérience, on a reçu l'estimateur $\hat{\alpha}$ de ce paramètre, qui, étant une variable aléatoire, ne coïncide générale-

ment pas avec la valeur α_0. Cependant l'écart $\Delta\alpha = \hat{\alpha} - \alpha_0$ de l'estimateur $\hat{\alpha}$ de la valeur réelle α_0 ne doit pas être grand ; si cet écart est excessif, on ne peut donc pas l'expliquer par des causes aléatoires et l'on est amené à conclure que l'hypothèse selon laquelle la valeur du paramètre est α_0 n'est pas confirmée et doit être rejetée. Ainsi donc, choisir le critère d'adéquation consiste à désigner la valeur critique de l'écart $\Delta\alpha_{cr}$ de telle sorte que la probabilité de dépasser cette valeur soit très petite. En qualité d'écart critique, on peut prendre les limites de l'intervalle de confiance à $(1-q) = 100\%$. On a alors la probabilité q pour que l'écart observé dépasse la valeur critique, c.-à-d. pour qu'une hypothèse correcte soit rejetée à tort. La grandeur q, appelée *niveau de signification statistique*, doit être choisie en fonction des conséquences auxquelles peut aboutir le rejet d'une hypothèse correcte. Pratiquement, on admet dans la plupart des cas $q = 0{,}05$ en prenant donc pour critère d'adéquation l'intervalle de confiance de 95 %. Or, dans les cas où l'on risque d'avoir des ennuis sérieux en rejetant à tort l'hypothèse correcte, on adopte $q = 0{,}01$, et même quelquefois moins.

Il est à noter que la méthode proposée, bien que permettant de réfuter avec une grande certitude les hypothèses fausses, ne peut garantir que l'hypothèse est vraie. Aussi, en voyant la valeur hypothétique α_0 tomber dans les limites de l'intervalle de confiance, ne retiendra-t-on pas l'hypothèse comme démontrée mais seulement comme conforme aux résultats de l'expérience. Pour savoir si l'hypothèse est correcte, on effectue des recherches spéciales.

PROBLÈMES AU CHAPITRE 5

5-1. Quels sont les événements S_1 et S_{15} de l'exemple 5-1?

5-2. Montrer que si l'espace des épreuves contient n éléments, il contient 2^n événements différents.

5-3. Les magasins d'une entreprise reçoivent des coupe-circuit provenant de deux usines dans le rapport 1 : 3. La première usine fournit 10 % de produits défectueux, la deuxième usine, 20 %. Appliquant la formule de la probabilité totale, chercher la probabilité pour qu'un coupe-circuit prélevé au hasard dans les magasins soit défectueux.

5-4. Construire le graphique de la fonction de répartition du nombre des points obtenus en jetant deux dés à jouer de l'exemple 5-1.

5-5. Chercher la probabilité pour que le temps d'attente du passager de l'exemple 5-5 ne soit pas supérieur à 3 mn.

5-6. Pour savoir combien il y a de poissons dans un étang, on tire 1 000 poissons, on les marque et on les remet à l'eau. Quel est le nombre de poissons dans l'étang pour avoir la probabilité maximale de rencontrer 10 poissons marqués parmi les 150 poissons tirés à nouveau?

5-7. Il s'agit de faire 100 biscuits aux raisins secs. Quelle est la quantité minimale de raisins secs qu'on doit mettre dans la pâte pour que la probabilité de n'avoir pas un seul raisin sec en aucun biscuit ne soit pas supérieure à 0,01?

DEUXIÈME PARTIE

OPTIMISATION DES PROCESSUS DE COMMANDE

CHAPITRE 6

STRUCTURE ET DESCRIPTION MATHÉMATIQUE DES PROBLÈMES DE COMMANDE OPTIMALE

6-1. TRAITS ESSENTIELS DU PROCESSUS DE COMMANDE

a) Notion de commande

Partout dans le monde ambiant (qu'il s'agisse de la nature, de la technique ou de la société humaine) on voit évoluer des processus dont le caractère dépend d'une foule de conditions accessoires et de nombreux facteurs. En changeant les conditions d'évolution de ces processus, l'homme peut faire varier leur caractère pour les adapter à ses objectifs. Cette intervention dans l'évolution naturelle d'un processus et le changement apporté à sa marche normale représentent l'essence de la commande. On peut donc dire que *la commande représente une organisation particulière de tel ou tel processus entreprise dans le but d'atteindre à coup sûr les objectifs fixés* [39].

Pour mieux comprendre l'essence du processus de commande, examinons le cas d'un chien qui poursuit un lièvre. Pour attraper le lièvre, le chien doit organiser ses actions d'une certaine manière, donc il doit les commander. Il s'ensuit que le processus de poursuite est un processus de commande.

Le début de la poursuite doit être précédé par l'apparition du lièvre, c.-à-d. par la création d'une situation qui engendre un but déterminé qu'il faut ou que l'on veut atteindre. Mais avant de se lancer à la poursuite du lièvre, le chien doit apprécier la situation qui s'est créée pour la confronter avec ses désirs et ses possibilités. L'appréciation de la situation aboutit à la prise d'une décision sur les actions à entreprendre: faut-il tâcher d'attraper le lièvre, ou non (le lièvre peut se trouver à une distance tellement grande que la poursuite soit privée de sens, le chien peut être fatigué, etc.)? Ce n'est qu'après la prise de la décision de poursuivre le lièvre que le chien procède à l'organisation de son mouvement ayant pour but d'attraper le lièvre en un temps minimal ou avec un minimum d'effort.

L'exemple cité met en évidence quatre étapes qui caractérisent tout processus de commande: l'apparition du but, l'appréciation de

la situation, la prise de décision et l'exécution de la décision prise. On remarquera que l'étape de l'apparition du but précède le début du processus de commande et c'est pour cette raison qu'elle ne sera pas examinée. D'autre part, compte tenu du fait que, lors de la commande des processus complexes, l'appréciation de la situation est réalisée sur la base d'une information collectée et dûment traitée, on arrive aux trois étapes suivantes du processus de commande:

1) collecte et traitement de l'information pour apprécier la situation créée;

2) prise de décision sur les actions les plus rationnelles à entreprendre;

3) exécution de la décision prise.

Parfois, il faut prévoir une quatrième étape incluant le contrôle de l'exécution de la décision.

Les différents types de problèmes de commande se distinguent justement par la méthode et par l'ordre d'exécution de ces opérations.

b) Types de problèmes de commande

Il y a un grand nombre de problèmes dont le mécanisme de collecte de l'information et d'exécution de la décision prise est tellement bien mis au point qu'il ne faut absolument pas penser à ces opérations lors de la réalisation de la commande. Dans ces problèmes, l'examen du processus de commande se réduit essentiellement à l'examen de la deuxième étape. Ces problèmes sont appelés *problèmes de décision à une étape.*

Mais, dans la majorité des cas, une telle approche n'est qu'une idéalisation et une simplification du processus réel de commande. En réalité, toutes les étapes du processus de commande sont étroitement liées et l'étape de la prise de décision implique un examen plus ou moins détaillé des procédés possibles de réalisation de la décision prise. Ainsi, pour prendre la décision de renoncer à la poursuite du lièvre, il faut se convaincre de l'inutilité de celle-ci, mais pour ce faire, il faut analyser, ne serait-ce que d'une manière approchée, les modes de poursuite possibles.

Parfois, dans des pareils cas, le processus de commande est décomposé en plusieurs étapes consécutives, la décision prise à une étape quelconque étant fonction des résultats obtenus par suite de l'exécution de la décision prise à l'étape précédente. Ces problèmes sont appelés problèmes de décision *à étapes multiples.* En qualité d'exemple on peut citer le processus de commande d'une fusée lancée de la Terre à la Lune. Ici, on distingue les étapes suivantes: placement de la fusée sur une orbite circumterrestre, organisation du mouvement de la fusée vers la Lune, transfert de la fusée sur une orbite circumlunaire, alunissage.

Dans cet exemple, les étapes isolées du processus de commande à étapes multiples ont été obtenues d'une façon bien naturelle. Mais dans beaucoup d'autres cas, la décomposition d'un processus de commande complexe en étapes distinctes avec mise en évidence de la commande de chacune de ces étapes constitue un problème bien ardu. Ainsi, au cours de la poursuite du lièvre, on a affaire à une situation qui varie sans cesse vu que le lièvre tâche d'échapper à la poursuite. Le chien doit apprécier continûment cette situation et prendre des décisions toujours nouvelles, conformément à la situation qui change en permanence, sans attendre les résultats finals de l'exécution des décisions précédentes. Dans les problèmes de ce type, on a affaire à des processus de commande *dynamiques continus.*

Ce qui vient d'être dit montre la complexité et la diversité des problèmes de commande. Mais on sous-estimerait les difficultés rencontrées lors de la résolution de ces problèmes si l'on ne tenait pas compte du fait que les processus de commande évoluent, en règle générale, dans un milieu ambiant complexe. L'évolution des processus de commande est influencée par différents facteurs extérieurs dont l'ensemble est souvent appelé *état de la nature.* Pour prendre une décision juste sur telle ou telle action, il faut estimer leurs résultats, mais, pour y parvenir, il faut connaître le caractère de la situation dans laquelle ces actions sont entreprises.

Toutefois, on peut considérer comme typiques les problèmes de commande où l'information dont on dispose est soit insuffisante pour permettre l'évaluation précise de la situation, soit dénaturée par des facteurs étrangers. Néanmoins, l'insuffisance de l'information ne nous dispense pas de prendre une décision. La particularité des problèmes de commande réside justement dans le fait que la décision doit être obligatoirement prise même si nous ne sommes pas en état d'évaluer avec précision les résultats auxquels aboutira la décision prise.

De cette façon, durant le processus de commande, il surgit le problème de décision dans les conditions où l'information sur la situation établie est soit insuffisante, soit dénaturée. Ce problème est dit *problème de décision en présence d'indétermination.*

c) Notion de recherche opérationnelle

Citons encore une classe spécifique de problèmes de commande liés à l'activité des grosses entreprises industrielles et qui peuvent être appelés *problèmes d'organisation-gestion* [40, 41].

Avant la révolution industrielle, la gestion d'une petite entreprise pouvait être assurée par une seule personne qui s'occupait des achats, de la planification et de la direction du travail, de l'écoulement des marchandises, de l'embauchage et du licenciement de personnel. L'envergure modeste de l'entreprise lui permettait de prendre des

décisions pratiques sans recourir aux méthodes scientifiques de gestion en s'appuyant seulement sur ses connaissances, expérience et intuition. Si quelques-unes des décisions prises ne s'avéraient pas les meilleures, elles n'entraînaient pas de pertes considérables, ou bien pouvaient être vite corrigées.

L'agrandissement des entreprises industrielles a rendu impossible la concentration des fonctions administratives entre les mains d'une seule personne. On a vu apparaître les chefs des services de la fabrication, des ventes, financiers, du personnel, etc. La mécanisation et l'automatisation de plus en plus poussées de la production ont déterminé la division ultérieure des fonctions administratives. Ainsi, les services de la fabrication se sont scindés en des groupes plus petits affectés à la résolution des problèmes concernant l'exploitation et l'entretien, le contrôle de la qualité, la planification, le ravitaillement, le stockage des produits finis, etc.

Chaque subdivision spécialisée d'une grosse entreprise effectue une partie bien déterminée de l'activité globale, étant guidée par les objectifs communs de l'entreprise. Mais en même temps, chaque subdivision spécialisée a ses propres objectifs. Tous ces objectifs ne vont pas toujours d'accord, quelquefois étant même contradictoires. En qualité d'exemple on peut examiner le problème des réserves de l'entreprise. Une certaine subdivision peut être intéressée par l'augmentation importante de ses réserves aux entrepôts afin de s'assurer une production sans à-coups. Mais si la capacité des entrepôts est limitée, les réserves des autres subdivisions doivent, dans ce cas, être réduites. Il en résulte un problème de type organisation-gestion qui consiste dans l'élaboration de la stratégie des réserves la plus avantageuse pour l'entreprise tout entière.

Lors de la résolution des pareils problèmes d'organisation-gestion, il faut faire preuve d'une compréhension subtile des objectifs des subdivisions isolées pour réaliser une concordance qui évite les contradictions aussi bien entre ces subdivisions qu'entre ces dernières et l'entreprise tout entière. Compte tenu des dommages importants qu'une grosse entreprise peut subir par suite de la prise des décisions non meilleures, il est évident qu'on ne peut pas résoudre ces problèmes en se basant exclusivement sur l'expérience personnelle et le bon sens, et qu'il faut recourir aux méthodes scientifiques.

L'élaboration des méthodes scientifiques de résolution des problèmes d'organisation-gestion fait l'objet d'une branche scientifique relativement jeune appelée *recherche opérationnelle*. Dans la terminologie de cette branche scientifique, le mot *opération* signifie une certaine mesure d'organisation dont la réalisation consiste à atteindre un certain but bien déterminé, par exemple la réglementation des réserves entreposées. Pour cela, il faut formuler les conditions qui caractérisent la situation dans laquelle sera réalisée la mesure, en particulier, les réserves nécessaires et les restrictions imposées par la

capacité des entrepôts dans l'exemple examiné. Le but de la recherche opérationnelle consiste à trouver et à donner une argumentation scientifique des procédés de réalisation de la mesure, procédés qui, dans un certain sens, sont les plus avantageux.

Une particularité spécifique des problèmes d'organisation-gestion consiste dans le fait que les conséquences de tel ou tel mode de résolution peuvent avoir des répercussions importantes sur le travail de toute l'entreprise. C'est pourquoi la prise de la décision finale revient toujours à une personne responsable (administrateur) pourvue de pouvoir, et sort des cadres de la recherche opérationnelle, qui n'a pour but que de munir cet administrateur de recommandations bien fondées sur la décision à prendre.

De cette façon, la recherche opérationnelle représente une branche scientifique qui s'occupe de l'élaboration des méthodes d'analyse des actions orientées vers un but déterminé (opérations) et de l'estimation comparée objective des décisions possibles. Bien que la recherche opérationnelle représente une branche scientifique autonome, qui a vu le jour au cours de la Seconde guerre mondiale, donc, avant l'apparition de la cybernétique, et qui à l'époque avait pour but la résolution des problèmes de défense antiaérienne de la Grande Bretagne, elle fait souvent appel à la cybernétique pour résoudre certains problèmes.

6-2. OPTIMISATION DU PROCESSUS DE COMMANDE

a) Critère de qualité de la commande

Dans ce qui suit, le problème de commande sera traité comme un problème mathématique, sans oublier qu'à la différence de beaucoup de problèmes mathématiques, il admet non pas une seule mais une multitude de solutions différentes [42]. Cela est conditionné par le fait que, dans les problèmes de commande, il y a d'habitude plusieurs procédés d'organisation de tel ou tel processus pour atteindre l'objectif fixé. Ainsi, lorsqu'il s'agit de la poursuite du lièvre, le chien peut organiser de façons différentes son mouvement, lors du lancement de la fusée vers la Lune, on peut choisir différentes trajectoires de vol. etc. C'est pourquoi le problème de commande pourrait être posé en tant que problème de recherche d'au moins un des procédés possibles menant à l'objectif fixé. Toutefois, une telle position de la question n'est d'habitude pas suffisante.

Si un problème quelconque admet plusieurs solutions, il surgit un problème supplémentaire visant à trouver parmi ces solutions celle qui est la meilleure d'un certain point de vue. On peut citer un grand nombre de problèmes de ce type. Ainsi, il y a beaucoup de procédés permettant de confectionner une boîte à partir d'une feuille de carton de dimensions données. Comme problème supplémentaire on peut se proposer l'obtention d'une boîte de capacité maximale.

D'une ville à une autre on peut voyager en empruntant différents moyens de transport (chemin de fer, avion, bateau, autobus, voiture automobile). Le problème supplémentaire consiste à choisir le moyen de transport le plus avantageux au point de vue temps de voyage, prix, confort, habitudes, etc. Une situation analogue se rencontre dans les problèmes de commande.

Si l'objectif de la commande peut être atteint en suivant plusieurs procédés différents, on peut imposer au procédé de commande des exigences supplémentaires dont le degré de satisfaction peut motiver la préférence que l'on donne à un certain procédé par rapport aux autres.

Souvent, la mise en œuvre du processus de commande est liée à la consommation de certaines ressources: temps, matières premières, combustible, énergie électrique. Donc, en choisissant le procédé de commande, il ne suffit pas de poser la question de l'atteinte de l'objectif, mais aussi celle des ressources engagées à cette fin. Dans ce cas, le problème de commande consiste à choisir dans l'ensemble des décisions, qui assurent l'atteinte de l'objectif, celle qui nécessite le minimum de ressources.

Dans d'autres cas, la préférence que l'on donne à un procédé de commande par rapport aux autres peut être motivée par d'autres exigences imposées au système de commande: le coût de l'entretien, la fiabilité, le degré d'approche de l'état atteint par le système par rapport à l'état requis, le degré de certitude des connaissances concernant l'état de la nature, etc.

L'expression mathématique de l'estimation quantitative du degré d'accomplissement des exigences imposées au procédé de commande est appelée *critère de qualité de la commande*. Le procédé de commande préféré ou procédé *optimal* sera celui pour lequel le critère de qualité de la commande atteindra son minimum (quelquefois son maximum). En choisissant, par exemple, le régime de vol, on peut prendre comme critère de qualité de la commande soit l'expression de la quantité de combustible consommé par unité de chemin parcouru, soit le chemin parcouru par unité de combustible consommé. Au régime à économie maximale, c.-à-d. au régime optimal, va correspondre soit la valeur minimale (premier cas), soit la valeur maximale (deuxième cas) du critère de qualité de la commande.

La définition de la commande optimale qui vient d'être donnée sera considérée comme préliminaire. Une définition plus rigoureuse sera donnée après l'examen des contraintes imposées au processus de commande.

b) Contraintes imposées au processus de commande

Le problème de commande optimale ou de commande en général ne se poserait pas si l'on n'imposait aucune contrainte au caractère d'évolution du système. Ainsi il n'y aurait pas de problème de

poursuite du lièvre si le chien pouvait parcourir en un clin d'œil la distance qui le sépare de son but. Il s'ensuit qu'en résolvant un problème de commande, on ne peut pas oublier que l'évolution de n'importe quel système est toujours liée à différentes contraintes.

Pour mieux se représenter les contraintes qui peuvent se rencontrer, examinons un exemple concret relatif à la conduite d'une voiture automobile. En réalisant la commande, le conducteur doit prendre en considération la puissance limitée de sa voiture, ce qui implique le transport d'une charge limitée sans dépassement d'une certaine vitesse limite. Grâce à l'inertie de la voiture, sa vitesse et sa direction de mouvement ne peuvent être changées qu'à une accélération limitée. Cela signifie que l'arrêt instantané et le changement instantané de direction sont exclus en cas d'apparition d'une situation dangereuse imprévue, ce qui limite, à son tour, la vitesse de déplacement. D'autre part, en choisissant son itinéraire, le conducteur doit tenir compte de la réserve de combustible dont il dispose et du ravitaillement à effectuer en route, etc.

Dans le cas général, il y a deux types de contraintes imposées au choix du procédé de commande [43]. Les contraintes du premier type sont constituées par les lois de la nature conformément auxquelles s'effectue l'évolution du système commandé. Lors de la formulation mathématique du problème de commande, ces contraintes se présentent d'habitude sous la forme d'équations algébriques, différentielles ou aux différences de l'objet commandé qui sont souvent appelées équations des liaisons. Les contraintes du deuxième type sont déterminées par la limitation des ressources utilisées pendant la commande ou d'autres grandeurs qui, en vertu des particularités physiques de tel ou tel système, ne peuvent ou ne doivent pas sortir de certaines limites. Au point de vue mathématique, les contraintes de ce type sont d'habitude exprimées sous la forme de systèmes d'équations algébriques ou d'inégalités qui lient les variables décrivant l'état du système.

c) Position du problème de commande optimale

Un problème de commande peut être considéré comme formulé mathématiquement si :

l'objectif de la commande, exprimé par le critère de qualité de la commande, est formulé ;

les contraintes du premier type, représentant un système d'équations différentielles ou aux différences qui limitent les modes possibles d'évolution du système, sont définies ;

les contraintes du deuxième type, représentant un système d'équations algébriques ou d'inégalités qui traduisent la limitation des ressources ou d'autres grandeurs utilisées dans la commande, sont définies.

Le mode de commande, qui satisfait à toutes les contraintes imposées et qui rend minimal (maximal) le critère de qualité de la commande, est appelé *commande optimale*.

6-3. DESCRIPTION MATHÉMATIQUE DE L'OBJET COMMANDÉ

a) Structure de l'objet commandé

Appelons l'objet commandé le système physique dont les processus sont soumis à notre commande. Il y a une grande diversité d'objets commandés dont la nature physique est très variée. Ainsi, on peut citer:

dispositifs techniques: voiture automobile, avion, fusée, tour, processus technologique, etc.;

entreprises industrielles: section, atelier, usine, branche industrielle;

systèmes économiques: économie de l'entreprise, économie de la branche industrielle, économie de l'état;

systèmes biologiques;

systèmes sociaux, etc.

Vu que les lois qui régissent les processus de commande sont communes quelle que soit la nature physique des objets commandés, on peut examiner la structure générale du processus de commande et donner sa description mathématique.

Désignons par x la variable qui définit l'état de l'objet commandé. Quelquefois, c'est une grandeur unidimensionnelle ou scalaire. En qualité d'exemples, on peut citer l'angle de rotation de l'arbre moteur, la vitesse de l'avion ou de la fusée, la pression de la vapeur dans la chaudière d'une machine à vapeur, la quantité d'objets aux entrepôts, le nombre d'avions basés sur un aérodrome, etc.

Toutefois, dans la majorité des cas, pour décrire l'objet commandé, il faut recourir à plusieurs variables $x_1, \ldots, x_N$:

lors de la description des systèmes mécaniques, les variables x_i représentent les coordonnées ou les vitesses des pièces en mouvement;

pour les systèmes électriques, les variables x_i seront des courants et des tensions;

s'il s'agit de l'économie, on peut avoir affaire aux capacités de production ou aux ressources de différentes branches industrielles;

pour un système biologique, les variables x_i peuvent caractériser la concentration des substances chimiques ou des médicaments dans divers organes.

Dans tous les cas susmentionnés, l'état de l'objet commandé est décrit par la variable multidimensionnelle, c.-à-d. vectorielle, x dont les composantes seront les variables x_i:

$$x = (x_1, \ldots, x_N). \tag{6-1}$$

Dans ce qui suit, la variable x sera appelée *variable* ou *vecteur* d'état de l'objet commandé.

Les variables x_i peuvent varier d'une façon continue dans une certaine gamme de valeurs ou bien elles peuvent prendre un ensemble fini de valeurs. Dans ce dernier cas, la variable x va prendre elle aussi un ensemble fini de valeurs et sa k-ième valeur sera notée par

$$x^{(k)} = (x_1^{(k)}, \ldots, x_N^{(k)}), \quad k = 1, \ldots, n. \qquad (6\text{-}2)$$

Alors, l'ensemble

$$X = \{x^{(1)}, \ldots, x^{(n)}\} \qquad (6\text{-}3)$$

représentera l'*espace des états possibles de l'objet commandé*. Parfois, l'espace X sera appelé *espace des décisions* pour souligner que le choix d'un certain $x \in X$ représente une solution possible du problème de commande.

Si les variables x_i peuvent varier d'une façon continue, c.-à-d. peuvent prendre une infinité de valeurs, l'espace des états possibles X du système sera un ensemble infini. Mais, dans ce cas aussi x_i ne peut d'habitude pas prendre des valeurs quelconques, car on peut lui imposer des contraintes qui, comme on l'a déjà remarqué, se présentent sous la forme de systèmes d'équations algébriques ou d'inégalités

$$f_i(x_1, \ldots, x_N) \{\leqslant, =, \geqslant\} 0, \quad i = 1, \ldots, m. \qquad (6\text{-}4)$$

Chacune des contraintes (6-4) conserve seulement un des signes, $\leqslant$, $=$ ou $\geqslant$, mais différentes contraintes peuvent avoir des signes différents. Les quantités m et N ne sont pas liées entre elles, de sorte que m peut être supérieure, inférieure ou égale à N. En particulier, m peut être égale à zéro, ce qui signifie que la contrainte (6-4) est absente. Souvent, quelques-unes ou même toutes les variables x_i satisfont à la condition de non-négativité

$$x_i \geqslant 0, \quad i = 1, \ldots, N. \qquad (6\text{-}5)$$

La condition de non-négativité des variables est très commode lors de la résolution numérique des équations qui décrivent le processus de commande. D'autre part, dans beaucoup de problèmes (économiques, par exemple) les variables x_i ne peuvent pas prendre des valeurs négatives de par leur sens physique (frais, production, marchandises transportées, capitaux placés de différentes manières, etc.). Pourtant, même les problèmes dont les variables peuvent avoir n'importe quel signe, c.-à-d. satisfont aux contraintes de la forme

$$x_i \geqslant a_i, \quad i = 1, \ldots N, \qquad (6\text{-}6)$$

où a_i sont des nombres arbitraires, peuvent facilement être transformés en problèmes à variables non négatives grâce à l'introduction

des nouvelles variables

$$y_i = x_i - a_i, \quad i = 1, \ldots, N. \tag{6-7}$$

Outre la variable x, que nous allons considérer comme étant une grandeur mesurable et contrôlable, l'état de l'objet commandé peut dépendre d'une foule de facteurs non contrôlables ou partiellement contrôlables déterminés par l'ensemble des conditions extérieures qui entourent l'objet commandé. Le pilote, par exemple, peut régler le régime de l'avion en faisant varier la hauteur et la vitesse de vol, qui, dans ce cas, sont des paramètres contrôlables. Mais la consommation de combustible est sensiblement influencée par les conditions atmosphériques extérieures que le pilote ne peut prendre en considération que partiellement, sans pouvoir les changer ni les prévoir d'une façon précise.

Désignons par ϑ l'ensemble entier des facteurs extérieurs incontrôlables qui influent sur le processus de commande et appelons cet ensemble *état de la nature.* Dans beaucoup de cas, en idéalisant quelque peu les phénomènes réels, on arrive à réduire la diversité infinie des conditions extérieures à un nombre fini d'états possibles de la nature $\vartheta^{(1)}, \ldots, \vartheta^{(l)}$, c.-à-d. à soumettre à l'examen l'ensemble fini

$$\Theta = \{\vartheta^{(1)}, \ldots, \vartheta^{(l)}\}, \tag{6-8}$$

appelé dans ce qui suit *espace des états de la nature*. Dans le cas général, les éléments $\vartheta^{(i)}$ de l'ensemble Θ sont des grandeurs multidimensionnelles

$$\vartheta^{(i)} = (\vartheta_1^{(i)}, \ldots, \vartheta_L^{(i)}), \quad i = 1, \ldots l. \tag{6-9}$$

L'impossibilité de contrôler parfaitement tous les facteurs extérieurs fait qu'au lieu de la connaissance précise de l'état de la nature ϑ on se limite souvent à la connaissance des probabilités $\xi(\vartheta)$ des différents états de la nature $\vartheta \in \Theta$. Les probabilités $\xi(\vartheta)$, obtenues d'une façon ou d'une autre pour tous les $\vartheta \in \Theta$ avant de procéder à la résolution du problème de commande, seront appelées *probabilités a priori* des états de la nature. Il est évident que les probabilités a priori $\xi(\vartheta)$, qui représentent la distribution des probabilités sur l'espace Θ, doivent satisfaire aux conditions (5-4).

Dans certains cas, l'état de la nature ϑ doit être considéré comme une variable aléatoire continue pour laquelle l'espace des états de la nature Θ se présente comme un ensemble infini. Dans ces cas, la distribution des probabilités $\xi(\vartheta)$ se transforme en densité de probabilité.

Pour décrire l'action orientée exercée sur l'objet commandé, introduisons la variable u et appelons-la *action de commande* ou tout simplement *commande*. De cette façon, le mot *commande* aura dans ce qui suit deux acceptions:

1) commande dans l'acception d'activité d'organisation développée dans le but d'atteindre certains objectifs;

2) commande dans l'acception d'action de commande, c.-à-d. d'une certaine grandeur physique que l'on fait varier suivant notre désir et qui exerce une influence sur le caractère des processus ayant lieu dans l'objet commandé en les orientant de la façon requise.

D'habitude, le contexte permet d'établir sans difficulté l'acception dans laquelle est utilisé le mot « commande ».

Lorsqu'on réalise la commande des objets complexes, il faut d'habitude faire appel à plusieurs actions de commande $u_1, \ldots, u_R$, de sorte que la commande u représente, dans le cas général, une grandeur multidimensionnelle

$$u = (u_1, \ldots, u_R). \qquad (6\text{-}10)$$

Ainsi, le pilote peut changer le régime de vol en utilisant la commande u qui comporte l'action exercée sur la quantité de combustible consommé (u_1) et l'action exercée sur le gouvernail de profondeur (u_2).

Dans les systèmes réels, les actions de commande u_i ne peuvent pas être prises indépendamment des différentes contraintes qui leur sont imposées. Désignons par U l'ensemble de toutes les valeurs des commandes u qui satisfont aux contraintes imposées. Avec cela, toute $u \in U$ sera une *commande admissible*. Par la suite, on rencontrera souvent des problèmes dans lesquels l'ensemble des commandes admissibles U est un ensemble fini

$$U = \{u^{(1)}, \ldots, u^{(r)}\}. \qquad (6\text{-}11)$$

Les valeurs prises par les variables d'état x permettent de juger de l'efficacité de la commande u adoptée dans le but d'atteindre les objectifs fixés. Mais, dans beaucoup de cas, cela s'avère incommode, car les objectifs de la commande exprimés à l'aide des variables d'état peuvent donner lieu à des relations assez compliquées. C'est pourquoi, outre les variables d'état x, il est commode d'introduire des *variables de sortie* q qui expriment les objectifs de la commande d'une façon explicite.

Parfois, on peut utiliser une variable d'état en qualité de variable de sortie, mais, dans le cas général, ce n'est pas ainsi. Par exemple, lors de la commande du régime de vol dans le but de minimiser la consommation de combustible sur un itinéraire donné, il est commode de prendre en qualité de variable de sortie le chemin parcouru par unité de combustible, tandis que les variables d'état seront représentées par la vitesse et l'altitude de vol de l'avion. Les variables de sortie peuvent être représentées aussi bien par des grandeurs physiques que par des indices économiques, par exemple, par le rendement, etc. En particulier, le critère de qualité de la commande peut aussi être examiné en tant que variable de sortie.

La variable de sortie q dépend tout d'abord de l'état x de l'objet commandé. Toutefois, elle peut être influencée par la commande adoptée u et par les facteurs extérieurs incontrôlables ϑ. Il s'ensuit que, dans le cas général, l'équation de la variable de sortie est de la forme

$$q = Q(x, u, \vartheta). \tag{6-12}$$

L'interaction de toutes les variables examinées ci-dessus est montrée par le schéma synoptique d'un objet commandé de la figure 6-1.

b) Equation d'évolution de l'objet commandé

Sous l'action des signaux de la commande u, l'état de l'objet commandé change. Le caractère des processus qui s'y produisent est défini par la vitesse de variation de la variable d'état de l'objet $\dot{x} = dx/dt$ qui, conformément à (6-1), est une grandeur multidimensionnelle

$$\dot{x} = (\dot{x}_1, \ldots, \dot{x}_N), \tag{6-13}$$

où $\dot{x}_1, \ldots, \dot{x}_N$ sont les vitesses de variation des composantes de la variable multidimensionnelle x.

Fig. 6-1. Schéma synoptique d'un objet commandé

Pour les systèmes dynamiques où les processus physiques évoluent continûment dans le temps, les vitesses $\dot{x}_i$ à un certain moment dépendent de l'état de l'objet commandé au même moment, état qui à son tour est défini par les valeurs de la variable d'état x, l'état de la nature ϑ et la commande adoptée u. Cette dépendance peut s'écrire sous la forme d'un système d'équations différentielles ordinaires

$$\dot{x}_i = g_i(x, u, \vartheta), \quad x_i(0) = c_i, \quad i = 1, \ldots, N, \tag{6-14}$$

où les grandeurs c_i ($i = 1, \ldots, N$) caractérisent l'état initial de l'objet commandé.

Parfois, il est commode de faire appel aux désignations vectorielles et de remplacer le système (6-14) par une seule équation

$$\dot{x} = g(x, u, \vartheta); \quad x(0) = c. \tag{6-15}$$

Ici, $c = (c_1, \ldots, c_N)$, tandis que par g on entend une fonction vectorielle

$$g = (g_1, \ldots, g_N). \tag{6-16}$$

L'introduction de la variable ϑ en tant qu'argument dans l'équation (6-15) s'avère incommode dans certains cas, car, durant l'évo-

lution d'un processus de commande, l'état de la nature reste souvent invariable. Toutefois, pour différents processus de commande, l'état de la nature peut être différent, ce qui se traduit par un changement de la forme revêtue par l'équation du mouvement. Il est commode de marquer cette dépendance entre la forme d'équation et l'état de la nature en ajoutant l'indice ϑ à la fonction g et d'écrire l'équation (6-15) comme suit:

$$\dot{x} = g_\vartheta(u, x), \quad x(0) = c. \tag{6-17}$$

Si l'état de la nature reste inchangé durant l'évolution de tous les processus de commande considérés, l'indice ϑ peut être omis.

6-4. CLASSIFICATION DES PROBLÈMES DE COMMANDE OPTIMALE

a) Problème de décision à une étape

D'habitude, dans les problèmes à une étape, on n'examine pas les méthodes de réalisation de la décision prise, ce qui signifie que l'on ne détermine pas la valeur et le caractère de l'action de commande u, mais on passe directement à la détermination de la valeur de la variable d'état x du système, valeur qui permet d'atteindre de la meilleure façon les objectifs de la commande.

Un problème de décision à une étape est considéré donné si l'on donne l'espace des états de la nature Θ avec la distribution de probabilités $\xi(\vartheta)$ pour tous les $\vartheta \in \Theta$, l'espace des décisions X et le critère de qualité de la décision adoptée qui, pour ce cas, sera appelé *fonction objectif*. Dans les ouvrages spécialisés, la fonction objectif est aussi appelée fonction de gains ou fonction de pertes. La fonction objectif qui exprime explicitement les objectifs de la commande peut être considérée en tant que grandeur de sortie de l'objet commandé et notée par q. La fonction objectif est une grandeur scalaire dépendant de l'état de la nature ϑ et de l'état de l'objet commandé x, qui, par analogie à (6-12), peut s'écrire sous la forme

$$q = Q(x, \vartheta). \tag{6-18}$$

On voit donc que le problème de décision à une étape représente un triplet

$$G = (X, \Theta, q), \tag{6-19}$$

où q est une fonction scalaire définie sur le produit direct des ensembles $X \times \Theta$. Trouver la solution de ce problème veut dire trouver un $x^* \in X$ tel que la fonction q atteigne son minimum, c.-à-d.

$$x^* = \{x \in X : Q(x, \vartheta) = \min\}. \tag{6-20}$$

Remarquons que s'il s'agit de maximiser la fonction q, cette opération se fait sans aucune difficulté, car si pour $x = x^*$, la fonction $Q(x, \vartheta)$ atteint son maximum, pour le même x, la fonction $-Q(x, \vartheta)$ atteindra son minimum.

Il y a plusieurs méthodes de résolution du problème de décision à une étape. L'application de telle ou telle méthode dépend de la façon dont on définit l'ensemble des décisions admissibles X, de l'information que l'on possède sur l'état de la nature et de la forme de la fonction objectif q. Ci-dessous on trouvera une brève caractéristique des méthodes principales.

Un problème est appelé *déterminé* s'il n'y a pas d'indétermination en ce qui concerne l'état de la nature. Dans les problèmes déterminés, l'espace des états de la nature Θ ne comporte qu'un seul élément ϑ_0 dont la probabilité est égale à l'unité. Dans ce cas, la fonction objectif dépendra seulement de l'état de l'objet commandé

$$q = Q(\vartheta_0, x) = q(x_1, \ldots, x_N). \tag{6-21}$$

Un problème déterminé à une étape est appelé *problème classique d'optimisation* [44] s'il contient des contraintes de la forme (6-4) parmi lesquelles il n'y a pas d'inégalités, ni de conditions de non-négativité ou de conditions demandant que les variables soient discrètes, si $m < N$ et si les fonctions $f_i(x_1, \ldots, x_N)$ et $q(x) = q(x_1, \ldots, x_N)$ sont continues et admettent des dérivées partielles du deuxième ordre au moins. Dans ce cas, le problème est formulé de la façon suivante. Soient les contraintes de la forme

$$f_i(x_1, \ldots, x_N) = 0, \quad i = 1, \ldots, m. \tag{6-22}$$

Trouver les valeurs de $x_1, \ldots, x_N$ vérifiant les équations (6-22) et rendant minimale la fonction $q(x_1, \ldots, x_N)$.

La particularité de ces problèmes consiste dans le fait qu'ils peuvent (au moins en principe) être résolus par des méthodes classiques basées sur le calcul différentiel. En effet, les équations (6-22) permettent d'éliminer m variables. De cette façon, la fonction objectif est ramenée à la forme

$$q(x_1, \ldots, x_N) = q_1(y_1, \ldots, y_{N-m}), \tag{6-23}$$

où par $y_1, \ldots, y_{N-m}$ sont désignées les variables restantes. Maintenant, le problème consiste à trouver les valeurs de $y_1, \ldots, y_{N-m}$ qui minimisent la fonction q_1 et auxquelles aucune contrainte n'est imposée. Cette solution peut être trouvée à partir du système d'équations obtenues en égalant à zéro les dérivées partielles de la fonction q_1:

$$\frac{\partial q_1}{\partial y_i} = 0, \quad i = 1, \ldots, N - m. \tag{6-24}$$

Mais cette voie est d'habitude liée à des calculs encombrants, ce qui rend nécessaires d'autres méthodes de résolution.

Ces derniers temps, on s'intéresse beaucoup aux méthodes de résolution des problèmes à une étape connues sous le nom de *programmation mathématique*. Ces méthodes permettent de trouver les valeurs des variables $x_1, \ldots, x_N$ satisfaisant aux contraintes de la forme (6-4), qu'elles soient exprimées par des égalités ou par des inégalités, et rendant minimale la fonction objectif $q(x_1, \ldots, x_N)$. Habituellement, on impose aux variables des conditions supplémentaires de non-négativité. Il faut remarquer que la programmation mathématique représente une forme algorithmique et non pas analytique de résolution du problème, c.-à-d. elle ne donne pas la formule du résultat final, mais indique seulement la procédure de calcul qui conduit à la solution du problème. Pour cette raison, l'efficacité des méthodes de la programmation mathématique est particulièrement évidente lors de l'utilisation des calculatrices numériques.

Le plus simple cas de problème de programmation mathématique est représenté par le problème de *programmation linéaire*. Ce dernier problème correspond au cas où les premiers membres des contraintes (6-4) et la fonction objectif (6-21) sont des fonctions linéaires de $x_1, \ldots, x_N$. Dans un problème de programmation linéaire, on demande de trouver les valeurs non négatives des variables $x_1, \ldots, x_N$ qui rendent minimale la fonction objectif

$$q(x_1, \ldots, x_N) = \sum_{j=1}^{N} b_j x_j \tag{6-25}$$

et satisfont au système de contraintes

$$\sum_{j=1}^{N} a_{ij} x_j \leqslant 0, \quad i = 1, \ldots, m. \tag{6-26}$$

Tout problème de programmation mathématique qui diffère du problème formulé ci-dessus est appelé *problème de programmation non linéaire*. Dans les problèmes de programmation non linéaire, la fonction objectif (6-21), ou les premiers membres des contraintes (6-4), ou bien les deux à la fois, sont des fonctions non linéaires de $x_1, \ldots, x_N$. Aux problèmes de programmation non linéaire se rapporte également le problème dans lequel la fonction objectif et les contraintes sont respectivement de la forme (6-25) et (6-26), mais on suppose, par exemple, que les variables soient des entiers. Ce problème est appelé problème de *programmation en nombres entiers*.

Les problèmes de programmation non linéaire sont beaucoup plus compliqués par rapport aux problèmes de programmation linéaire et à l'heure actuelle on ne dispose de méthodes de calcul que pour un nombre restreint des problèmes de ce type. Ci-après, on trouvera un exposé détaillé des méthodes de résolution des problèmes de programmation linéaire. Les méthodes de résolution des problèmes de programmation non linéaire peuvent être trouvées dans [10, 44 à 46].

Un problème de décision à une étape est appelé *stochastique* si l'espace des états de la nature Θ comporte plus d'un élément, de sorte que l'on ne connaît pas l'état réel de la nature ϑ, mais la distribution de probabilités $\xi(\vartheta)$ sur l'espace Θ.

Les problèmes stochastiques où il faut trouver les valeurs des variables satisfaisant aux contraintes (6-4) et minimisant la fonction objectif (6-18) sont des problèmes de programmation stochastique. Toutefois, dans beaucoup de cas, on parvient à ramener les problèmes de programmation stochastique aux problèmes de programmation linéaire ou non linéaire en définissant d'une façon un peu différente la fonction objectif. En effet, étant donné que l'état de la nature ϑ est une variable aléatoire avec la distribution de probabilités $\xi(\vartheta)$ sur l'espace Θ, la valeur de $Q(\vartheta, x)$, pour $x = (x_1, \ldots, x_N)$ donné, sera aussi une variable aléatoire avec la même distribution de probabilités $\xi(\vartheta)$ sur l'espace Θ. Pour cette raison, dans ce cas, il est rationnel de prendre en tant que fonction objectif la valeur moyenne de la fonction $Q(\vartheta, x)$ sur l'espace Θ.

Ainsi, pour les problèmes stochastiques, la fonction objectif peut être définie conformément à (5-64) par l'expression

$$q_1(x) = \sum_{\vartheta \in \Theta} \xi(\vartheta) Q(\vartheta, x). \qquad (6\text{-}27)$$

Vu que $q_1(x)$ est une fonction déterminée de x, le problème qui consiste à trouver les variables $x_1, \ldots, x_N$ satisfaisant aux contraintes (6-4) et minimisant la fonction objectif (6-27) peut être résolu à l'aide des méthodes de la programmation linéaire ou non linéaire.

Un cas important de problème de décision stochastique à une étape se présente lorsque les grandeurs x_i $(i = 1, \ldots, N)$ ne peuvent prendre qu'un ensemble fini de valeurs, c.-à-d. les contraintes imposées aux valeurs des variables sont données sous la forme d'un espace des décisions X défini par la relation (6-3). Les méthodes de résolution des problèmes de ce type font l'objet de la branche des mathématiques appelée *théorie des décisions statistiques.*

Actuellement, on s'intéresse beaucoup aux problèmes où la décision n'est pas prise par une seule personne mais par plusieurs (par exemple, deux) dont les intérêts sont opposés. Comme exemple on peut citer le problème de la poursuite dans lequel la distance qui sépare le poursuiveur du poursuivi dépend des décisions de ces deux personnes. Dans ce cas, le poursuiveur tâche de réduire cette distance au minimum, tandis que le poursuivi tend à la rendre maximale. Des problèmes pareils ont été appelés *situations de conflit,* les méthodes de leur résolution étant examinées par la *théorie des jeux.* Les personnes qui participent au jeu s'appellent *joueurs.*

Etant donné que dans une situation de conflit les décisions sont prises par chacun des joueurs indépendamment des décisions prises par l'autre joueur, en décrivant mathématiquement la situation de

conflit il convient de considérer l'espace des décisions comme un produit direct de deux ensembles

$$X \times Y, \tag{6-28}$$

où $X = \{x_1, \ldots, x_n\}$ est l'espace des décisions du premier joueur, $Y = \{y_1, \ldots, y_m\}$ l'espace des décisions du deuxième joueur.

Les éléments de l'espace des décisions $X \times Y$ seront des couples de la forme (x, y), $x \in X$, $y \in Y$, c.-à-d. seront définis par les décisions prises aussi bien par le premier joueur que par le deuxième. Pour plus de simplicité, admettons qu'il n'y a pas d'indétermination dans l'état de la nature. Alors, la fonction objectif ne dépendra que des éléments de l'espace $X \times Y$ et aura la forme

$$q = Q(x, y). \tag{6-29}$$

L'opposition des intérêts des joueurs consiste dans le fait que le premier joueur, qui fait son choix dans l'ensemble X, tâche de minimiser la fonction objectif, tandis que le deuxième joueur, qui a à sa disposition l'ensemble Y, tend à la maximiser. Ainsi, l'essence d'une situation de conflit réside en ce que chaque joueur doit prendre la meilleure décision à son point de vue, compte tenu du fait que son adversaire en fera autant.

b) Problèmes dynamiques d'optimisation de la commande

Parmi la grande variété de problèmes de la cybernétique, une place importante revient aux problèmes dont l'objet commandé subit une évolution perpétuelle sous l'action de divers facteurs extérieurs et intérieurs. Les problèmes de commande de pareils objets se rapportent à la classe des *problèmes de commande dynamiques*.

Un objet est appelé commandable (gouvernable) si le caractère de son évolution peut être changé en actionnant sur certains des facteurs à l'influence desquels il est soumis. Comme il a été déjà indiqué, les actions orientées vers un but bien déterminé sont appelées commandes et désignées par $u(t)$.

Le caractère de l'évolution de l'objet commandé est défini par le système d'équations différentielles (6-14) qu'il est commode d'écrire de façon condensée sous la forme vectorielle comme une seule équation différentielle (6-17). La commande $u(t)$ fait partie de l'équation (6-17), donc cette dernière ne définit pas une évolution concrète de l'objet, mais seulement ses possibilités techniques qui peuvent être mises en valeur en adoptant telle ou telle commande comprise dans l'espace des commandes admissibles U.

On peut porter une estimation sur l'efficacité de tel ou tel mode de commande en introduisant une fonction objectif du type (6-12)

qu'il est commode d'écrire, dans ce cas, sous la forme

$$q = Q_\vartheta [x(t), u(t)]. \qquad (6\text{-}30)$$

De cette façon, si $u(t)$ est la consommation instantanée de combustible et si $x(t)$ est la vitesse instantanée de l'avion, alors, du point de vue de la consommation de combustible, la qualité de la commande peut être caractérisée à tout moment par la grandeur $q(t) = u(t)/x(t)$ (consommation instantanée de combustible par unité de chemin parcouru) qui évidemment sera fonction de l'état de la nature ϑ, c.-à-d. de l'ensemble des facteurs extérieurs qui déterminent les conditions de vol.

La fonction objectif de la forme (6-30) n'est utilisée que rarement, car elle donne seulement l'estimation des valeurs instantanées du processus commandé, tandis que dans la grande majorité des problèmes il est nécessaire d'estimer les processus qui évoluent dans l'objet commandé durant toute la période de commande allant de 0 à T.

Dans beaucoup de cas, la fonction odjectif peut être choisie de façon que l'on arrive à estimer le processus qui se déroule dans l'objet commandé en intégrant la fonction objectif sur toute la période de commande, c.-à-d. on peut prendre en tant que critère de qualité de la commande la fonctionnelle

$$J(u) = \int_0^T Q_\vartheta [x(t), u(t)]\, dt. \qquad (6\text{-}31)$$

Ainsi, lorsque le sens physique de la fonction objectif se traduit par des pertes, l'expression (6-31) définit les pertes totales durant tout le processus de commande.

Parfois, on arrive à donner en tant qu'objectif de la commande la marche désirée du processus $z(t)$. Ici, on peut prendre en qualité de fonction objectif le carré ou la valeur absolue de l'écart du processus $x(t)$ par rapport à la marche désirée

$$q = [x(t) - z(t)]^2, \quad q = |x(t) - z(t)|. \qquad (6\text{-}32)$$

Dans ces cas, le critère de qualité de la commande (6-31) définit l'erreur totale quadratique ou l'erreur absolue.

Dans les problèmes de commande dynamiques, outre les contraintes de la forme (6-11), qui définissent l'espace des commandes admissibles U, on a affaire à des contraintes intégrales de la forme

$$\int_0^T H_\vartheta [x(t), u(t)]\, dt \leqslant K = \text{const}. \qquad (6\text{-}33)$$

Très souvent, par exemple, on est obligé de réduire les limites de variation des valeurs instantanées d'un certain paramètre $a(x, u)$

durant le processus de commande. Désignons par a_0 la valeur du paramètre a qu'il n'est pas recommandable de dépasser. Si la fonction figurant sous le signe d'intégration $H(x, u)$ est définie à partir de la relation

$$H(x, u) = \begin{cases} 0 & \text{pour} \quad a(x, u) \leqslant a_0; \\ [a(x, u) - a_0]^2 & \text{pour} \quad a(x, u) > a_0, \end{cases} \tag{6-34}$$

la contrainte intégrale (6-33) exprimera l'exigence imposant que la valeur instantanée du paramètre a ne dépasse a_0 que pour un court laps de temps et d'une quantité insignifiante. Cette condition sera d'autant plus rigoureuse que plus petit sera K. Ainsi, pour $K = 0$, la contrainte (6-33) n'admettra pas que a dépasse a_0 si insignifiant que soit ce dépassement.

Des contraintes de la forme (6-33) surgissent aussi lorsqu'on a affaire à des ressources limitées: c'est l'énergie ou le combustible disponibles qui peuvent être limités s'il s'agit de la trajectoire, etc.

Les relations citées permettent de formuler de la façon suivante la définition de la commande optimale des systèmes dynamiques. Une commande $u^*(t)$ choisie dans l'espace des commandes admissibles U est appelée *optimale* pour l'objet décrit par l'équation différentielle (6-17) si elle minimise le critère de qualité (6-31) en présence des contraintes données imposées aux ressources utilisées (6-33).

Les problèmes de commande dynamiques de même que les problèmes à une étape peuvent être *déterminés* si l'espace des états de la nature Θ ne comprend qu'un seul élément $\vartheta = \vartheta_0$, ou *stochastiques* lorsque l'espace des états de la nature Θ comporte plus d'un élément et lorsqu'on donne la distribution a priori des probabilités $\xi(\vartheta)$ sur l'espace Θ.

Parmi les problèmes stochastiques une place importante est occupée par les problèmes de commande *adaptative*. Cette commande est adoptée dans les cas où les données a priori sur l'état de la nature sont insuffisantes pour assurer une commande efficace, ou bien si l'on n'est pas en possession d'une description mathématique suffisamment précise de l'objet commandé lui-même. La commande adaptative a pour but de préciser les données concernant l'état de la nature ou les propriétés de l'objet commandé directement durant le processus de commande de ce dernier en essayant différents modes de commande pour trouver celui qui, dans les conditions concrètes données, est le plus efficace.

c) Commande de l'état final

Il y a toute une série de cas où le caractère de l'évolution de l'objet durant le processus de commande ne présente pas d'intérêt, car ce n'est que l'état auquel arrivera l'objet à la fin du processus

de commande qui nous intéresse. En voici quelques exemples: l'atterrissage aux instruments qui impose des exigences très rigoureuses à la vitesse et à la position que l'avion occupe à l'instant de contact du sol, la livraison de la charge dans le délai fixé au point de destination, l'obtention à la fin de l'année de la productivité du travail fixée, etc. Ces problèmes ont été appelés *problèmes de commande de l'état final.*

Désignons par $x(T)$ l'état de l'objet à l'instant final. La fonction objectif pour ce problème sera de la forme:

$$q = Q_0 [x(T)]. \tag{6-35}$$

Etant donné que $x(T)$ dépend du caractère de la commande adoptée $u(t)$, la valeur de q sera aussi fonction de cette commande. C'est pourquoi, pour ce cas, le problème du choix de la commande optimale peut être formulé de la façon suivante: choisir dans l'espace des commandes admissibles U une commande $u^*(t)$ telle que la fonction objectif (6-35) soit minimisée pour l'objet décrit par l'équation différentielle (6-17) en présence des contraintes (6-33) imposées aux ressources utilisées.

d) Jeux différentiels

La théorie des jeux différentiels c'est l'extension de la théorie des jeux pour les problèmes à une étape au cas des problèmes de commande dynamiques. A un jeu différentiel comme à un jeu ordinaire participent deux personnes appelées joueurs dont les intérêts sont opposés. D'autre part, chacun des joueurs peut, dans une certaine mesure, influer sur la situation en actionnant sur l'objet par l'intermédiaire de sa propre commande. Cela signifie que dans les jeux différentiels l'espace des commandes admissibles représente le produit direct des ensembles

$$U \times V, \tag{6-36}$$

où U est l'espace des commandes admissibles du premier joueur contenant les éléments $u(t)$ et V l'espace des commandes admissibles du deuxième joueur à éléments $v(t)$. De cette façon, les éléments de l'espace des commandes admissibles seront constitués par des couples de la forme (u, v), $u \in U$, $v \in V$.

S'il n'y a pas d'indétermination relative à l'état de la nature, l'équation différentielle qui décrit l'évolution de l'objet commandé sera obtenue à partir de (6-17) par la substitution du couple (u, v) à u, c.-à-d. sera de la forme

$$\dot{x} = g(x, u, v), \quad x(0) = c. \tag{6-37}$$

Habituellement, les jeux différentiels représentent des problèmes de commande de la valeur finale. Pour cette raison, la fonction objectif dépend de l'état de l'objet à l'instant final T, état qui à son tour est défini par les commandes utilisées et peut donc s'écrire

$$q = Q\,[x\,(T)],\quad x\,(T) = f\,[u\,(t),\ v\,(t)]. \tag{6-38}$$

Les buts des joueurs sont opposés, car pendant que le premier joueur, en choisissant la commande u dans l'espace U, tend à minimiser la fonction objectif, les intérêts du deuxième joueur imposent qu'il choisisse dans l'espace V une commande v telle que la fonction objectif atteigne son maximum. La façon de trouver, dans cette situation contradictoire, la meilleure commande pour chacun des joueurs constitue l'essence de la théorie des jeux différentiels [47].

Voici quelques exemples de situations typiques pour les jeux différentiels: les combats, les combats aériens, le football, la poursuite d'un navire par une torpille, la défense des objectifs contre une attaque ennemie, etc.

6-5. PROCESSUS DE COMMANDE À ÉTAPES MULTIPLES

a) Comportement d'un système dynamique en tant que fonction de l'état initial

Dans beaucoup de cas, on arrive à simplifier sensiblement la recherche de la commande optimale dans les systèmes dynamiques si l'on réussit à décomposer, d'une façon naturelle ou artificielle, le processus de commande en étapes isolées.

Pour entreprendre l'examen sous une forme générale, nous allons considérer que l'état de l'objet est décrit par la variable multidimensionnelle

$$x = (x^{(1)}, \ldots, x^{(n)}). \tag{6-39}$$

En supposant que le processus soit non commandable et qu'il n'y ait pas d'indétermination dans l'état de la nature, écrivons par analogie à (6-17) l'équation différentielle, qui définit l'évolution de l'objet, sous la forme

$$\dot{x} = g\,(x),\quad x\,(0) = c. \tag{6-40}$$

La solution de cette équation se met d'habitude sous la forme $x = x\,(t)$ en soulignant de cette façon qu'elle dépend du temps. Toutefois, il n'est pas moins important que la solution de l'équation (6-40) dépende de l'état initial c. C'est pourquoi une écriture plus rigoureuse montre la dépendance explicite de la solution x aussi bien du temps que de l'état initial:

$$x = x\,(c,\ t) = x\,[x\,(0),\ t]. \tag{6-41}$$

Une telle écriture permet de considérer l'état du système à un moment arbitraire t comme une transformation de l'état initial $x(0) = c$ sur l'intervalle t.

Examinons l'évolution de l'objet sur l'intervalle de 0 à t_2 que nous allons décomposer, à l'aide d'un point intermédiaire t_1, en deux intervalles de durée t_1 et $\tau = t_2 - t_1$ comme le montre la figure 6-2. Examinons trois états de l'objet commandé:

état initial $x(0) = c$;

état $x(c, t_1)$ correspondant au moment intermédiaire t_1;

état $x(c, t_2)$ correspondant au moment final t_2.

Fig. 6-2. Décomposition d'un intervalle en deux étapes

La description du dernier état peut être abordée de deux manières. Cet état peut être considéré comme une transformation de l'état initial $x(0) = c$ sur l'intervalle $t_2 = t_1 + \tau$

$$x(c, t_2) = x(c, t_1 + \tau), \tag{6-42}$$

ou bien en tant qu'une transformation de l'état $x(c, t_1)$ sur l'intervalle τ

$$x(c, t_2) = x[x(c, t_1), \tau]. \tag{6-43}$$

Etant donné que ces deux expressions traduisent le même état, en les égalant on obtient la relation

$$x(c, t_1 + \tau) = x[x(c, t_1), \tau]. \tag{6-44}$$

b) Représentation d'un processus dynamique sous la forme d'une suite de transformations

Supposons que le processus dynamique $x(c, t)$ puisse être présenté sur l'intervalle allant de 0 à t' d'une façon naturelle ou artificielle sous la forme d'un processus à étapes multiples et proposons-

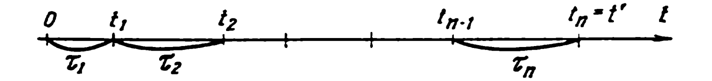


Fig. 6-3. Décomposition d'un intervalle en n étapes

nous de le décrire d'une façon convenable. Pour obtenir un processus à étapes multiples, l'intervalle de 0 à t' doit être décomposé en n étapes successives de durées $\tau_1, \tau_2, \ldots, \tau_n$, comme il est montré sur la figure 6-3. Désignons par t_k ($k = 0, \ldots, n$) les moments où prend fin la k-ième étape de sorte que $t_{k+1} = t_k + \tau_{k+1}$, et par

x_k l'état de l'objet au moment t_k:

$$x_k = x(c, t_k). \tag{6-45}$$

L'état

$$x_{k+1} = x(c, t_{k+1}) = x(c, t_k + \tau_{k+1}). \tag{6-46}$$

Suivant (6-44) et (6-45), cette expression peut être mise sous la forme

$$x_{k+1} = x[x(c, t_k), \tau_{k+1}] = x(x_k, \tau_{k+1}). \tag{6-47}$$

La relation (6-47) représente l'état de l'objet x_{k+1} en tant que résultat de la transformation de l'état x_k subie durant la $(k+1)$-ième étape.

Introduisons l'opérateur T pour désigner la transformation de l'état du processus durant une étape:

$$T(x_k) = x(x_k, \tau_{k+1}), \quad k = 0, \ldots, n-1. \tag{6-48}$$

Alors la relation (6-47) va s'écrire

$$x_{k+1} = T(x_k). \tag{6-49}$$

En mettant $k = 0, 1, \ldots, n-1$, on peut décrire le processus dynamique tout entier sous la forme d'une suite de transformations

$$x_0 = c, \quad x_1 = T(x_0), \ldots, x_n = T(x_{n-1}). \tag{6-50}$$

c) Processus de commande à étapes multiples

Le processus dynamique décrit par la transformation (6-49) est un processus non commandable. Pour obtenir un processus commandable à étapes multiples, il faut avoir la possibilité d'effectuer à chaque étape non pas une transformation $T(x_k)$ mais une transformation comprise dans l'ensemble des transformations $T_1(x_k), \ldots, T_r(x_k)$.

Il est commode de considérer qu'une forme de transformation concrète va dépendre d'un paramètre u_k qui, à la k-ième étape, peut prendre une valeur de l'ensemble de valeurs U_k. Dans ce qui suit, le paramètre u_k sera appelé commande, tandis que l'ensemble U_k sera l'espace des commandes admissibles à la k-ieme étape. Maintenant, la transformation réalisée à la k-ième étape peut s'écrire sous la forme

$$x_{k+1} = T(x_k, u_k), \quad u_k \in U_k. \tag{6-51}$$

Si l'on met successivement dans la relation (6-51) $k = 0, 1, \ldots, n-1$ et si l'on tient compte de l'état initial x_0, on obtient la description de tout le processus commandable à étapes multiples:

$$x_{k+1} = T(x_k, u_k), \quad u_k \in U_k;$$
$$k = 0, 1, \ldots, n-1; \quad x_0 = x(0) = c. \tag{6-52}$$

La relation (6-52), qui est appelée équation aux différences de l'objet commandé, est analogue à l'équation différentielle (6-17) qui décrit un processus dynamique continu.

Exemple 6-1. Supposons que la variable d'état soit une variable bidimensionnelle $x = (x^{(1)}, x^{(2)})$ qui peut prendre les valeurs déterminées géométriquement par les mailles de la grille représentée sur la figure 6-4, *a*. Pour passer d'une maille de grille à la maille suivante, on peut utiliser à chaque étape une

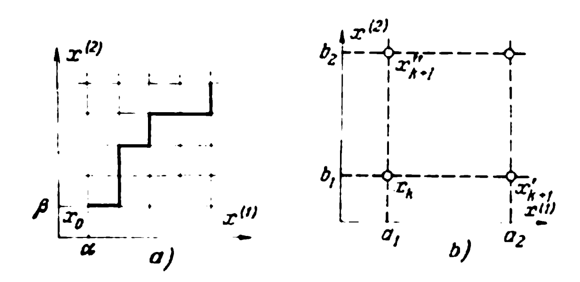


Fig. 6-4. Processus à étapes multiples à une variable bidimensionnelle

des deux commandes possibles: $u_k = 0$ (déplacement horizontal) ou $u_k = 1$ (déplacement vertical). Il s'ensuit que l'espace des commandes admissibles est le même pour toute étape et est égal à $U_k = \{0, 1\}$, $k = 0, 1, \ldots, n-1$.

Examinons une case de la grille donnée représentée sur la figure 6-4, *b*, au coin inférieur gauche de laquelle le système est arrivé après la k-ième étape, de façon que $x_k = (a_1, b_1)$. La valeur de $x_{k+1} = T(x_k, u_k)$ est fonction de la commande adoptée. La figure 6-4, *b* montre que les relations ci-dessous ont lieu:

$$T[(a_1, b_1), 0] = (a_2, b_1) = x'_{k+1};$$

$$T[(a_1, b_1), 1] = (a_1, b_2) = x''_{k+1}.$$

On peut décrire une trajectoire concrète du mouvement du système en indiquant l'état initial x_0 et la suite des commandes adoptées. Ainsi, la trajectoire marquée en trait fort sur la figure 6-4, *a* s'obtient en adoptant la commande $u = (01101001)$, les conditions initiales étant $x_0 = (\alpha, \beta)$.

d) Critère de qualité de la commande pour un processus à étapes multiples

Nous allons considérer que la qualité de la commande est déterminée par la valeur de la fonction objectif q dont la valeur numérique peut être considérée en tant que pertes subies en adoptant telle ou telle commande. Les pertes correspondant à une étape seront fonction de l'état du processus au début de cette étape et de la commande qui y est adoptée, c.-à-d.

$$q_k = Q(x_k, u_k), \quad u_k \in U_k. \tag{6-53}$$

Comme critère de qualité de la commande on peut prendre les pertes totales correspondant à toutes les n étapes du processus en présentant alors le critère de qualité de la commande d'un processus à n étapes sous la forme

$$J_n(u) = \sum_{k=0}^{n-1} Q(x_k, u_k). \tag{6-54}$$

Ici, on désigne par u la suite de commandes, c.-à-d. un ensemble ordonné de la forme

$$u = (u_0, \ldots, u_{n-1}), \quad u_0 \in U_0, \ldots, u_{n-1} \in U_{n-1}. \tag{6-55}$$

Si l'on considère un processus à étapes multiples de commande de la valeur finale, les pertes vont dépendre seulement de l'état de l'objet commandé à la fin du processus, état qui à son tour est fonction de l'état initial et des commandes adoptées à chaque étape:

$$q = q(x_n) = Q(x_0, u), \tag{6-56}$$

où u est définie par la relation (6-55). La grandeur scalaire définie par l'expression (6-56) sera appelée *fonction objectif* du processus de commande à étapes multiples.

Le problème de recherche de la commande optimale pour un processus à étapes multiples peut être formulé de la façon suivante. Pour un système dynamique dont les processus sont décrits par l'équation aux différences (6-52), il faut trouver une suite de commandes $u_0, u_1, \ldots, u_{n-1}$, satisfaisant aux contraintes de la forme $u_k \in U_k$, $k = 0, 1, \ldots, n-1$, telle que le critère de qualité de la commande (6-54) ou la fonction objectif (6-56) atteigne son minimum.

Etant donné que les processus de commande à étapes multiples représentent un cas particulier des processus de commande dynamiques, on peut y trouver tous les types de problèmes examinés déjà dans les problèmes de commande dynamiques. Ainsi, lorsque l'espace des états de la nature Θ ne comporte qu'un seul élément ϑ_0, le problème de commande à étapes multiples est appelé *déterminé*. Dans le cas contraire, il se rapporte à la classe de problèmes *stochastiques*.

Il est à remarquer que l'on n'arrive pas toujours à situer d'une façon univoque tel ou tel problème dans la classe des problèmes à étapes multiples ou dans celle des problèmes à une étape. Ainsi, lorsque dans un problème à étapes multiples on considère la suite des valeurs de la variable, prise aux étapes isolées, en tant qu'une variable multidimensionnelle, le problème à étapes multiples se transforme en un problème à une étape.

6-6. PROBLÈME DÉTERMINÉ D'OPTIMISATION A UNE ÉTAPE A UNE VARIABLE D'ÉTAT UNIDIMENSIONNELLE

a) Position du problème

Parmi la grande variété de problèmes d'optimisation de la commande, le problème à une étape à une variable unidimensionnelle occupe une place un peu particulière. Cela s'explique par le fait que, d'une part, les problèmes de ce genre se rencontrent très souvent dans la pratique et, d'autre part, beaucoup de problèmes à étapes multiples et de problèmes à une étape à une variable multidimensionnelle comportent en tant qu'étapes isolées de leur résolution le problème à une étape à une variable unidimensionnelle. Bien que le problème d'optimisation à une étape à une variable unidimensionnelle puisse, dans la majorité des cas, être résolu par les moyens de l'analyse mathématique classique étudiée dans les mathématiques supérieures, en mettant au clair certaines propriétés de la solution et en utilisant certains procédés de calcul, on arrive souvent à simplifier considérablement la marche de la résolution.

Le problème déterminé d'optimisation à une étape à une variable unidimensionnelle peut être formulé de la façon suivante. Soient R l'ensemble des nombres réels, $X \subseteq R$ l'ensemble des valeurs admissibles de la variable unidimensionnelle x, $q(x)$ une fonction de x définie sur l'ensemble X. On demande de trouver la valeur de $x \in X$ (notée dans ce qui suit par x^*) qui minimise ou maximise $q(x)$, c.-à-d. satisfait à la condition

$$x^* = \{x \in X: q(x) = \min (\max)\}. \qquad (6\text{-}57)$$

Les méthodes de résolution de ce problème dépendent du caractère de l'ensemble X des valeurs admissibles de la variable, que nous allons supposer défini sur un intervalle réel fini. Ci-après seront examinés quelques cas particuliers.

b) Cas d'un ensemble fini de solutions admissibles

Examinons le cas où l'ensemble des solutions admissibles est un ensemble fini:

$$X = \{x_1, \ldots, x_N\}. \qquad (6\text{-}58)$$

Dans ce cas, la fonction objectif $q(x)$ sera représentée par l'ensemble des nombres réels Q qui peuvent être considérés comme une application de l'ensemble X dans l'ensemble des nombres réels R:

$$Q: X \to R. \qquad (6\text{-}59)$$

Les éléments de l'ensemble Q, c.-à-d. les valeurs de $q(x)$ pour tous les $x \in X$, peuvent être calculés.

La voie générale qu'il faut suivre pour trouver la valeur optimale x^* consiste à comparer entre eux deux à deux tous les éléments de l'ensemble Q et à trouver ainsi l'élément minimal ou maximal. Cette voie est simple à condition que l'ensemble X, donc l'ensemble Q aussi, possède un nombre réduit d'éléments. Mais si l'ensemble X comporte un grand nombre d'éléments, cette méthode exige beaucoup de temps, ce qui est lié avant tout à la nécessité de calculer à maintes reprises les valeurs de $q(x)$.

Cette procédure ne peut être simplifiée que par la diminution du nombre d'éléments de l'ensemble X. Cela peut se faire en éliminant les éléments qui à coup sûr ne contiennent pas la valeur x^* (cette certitude est acquise non pas à la suite des calculs des valeurs de $q(x)$ mais sur la base de certains autres indices accessoires). Dans ce qui suit, on donnera des exemples d'utilisation de cette méthode qui, bien sûr, ne doit pas être exclue lors de la résolution des problèmes aux variables multidimensionnelles.

c) Cas d'un ensemble infini borné des solutions admissibles

Le plus souvent, on rencontre le cas où l'ensemble X des valeurs admissibles de la variable représente un intervalle réel continu borné par les nombres a et b:

$$X = \{x \in R : a \leqslant x \leqslant b\} = [a, b]. \tag{6-60}$$

Supposons que la fonction $q(x)$ soit continue et différentiable en chaque point de l'intervalle $[a, b]$ à l'exception, peut-être, d'un nombre fini de points. Une fonction satisfaisant à ces conditions est représentée sur la figure 6-5.

Fig. 6-5. Exemple d'une fonction objectif

Désignons comme auparavant par x^* la valeur de $x \in X$ pour laquelle la fonction $q(x)$ atteint son minimum ou son maximum. Pour déterminer x^*, utilisons le théorème bien connu de l'analyse mathématique suivant lequel une fonction continue et différentiable ne peut présenter un maximum ou un minimum qu'aux points où sa dérivée est nulle. Ces points peuvent être trouvés de l'équation

$$q'(x) = 0. \tag{6-61}$$

Mais l'intervalle $[a, b]$ peut contenir des points où il n'y a pas de dérivée, comme, par exemple, le point M sur la figure 6-5 ou bien les extrémités a et b de l'intervalle. En ces points, qui ne véri-

fient pas la condition (6-61), la fonction $q(x)$ peut aussi présenter un maximum ou un minimum.

Pour formuler la règle générale, introduisons les désignations suivantes. Désignons par

$$S_1 = \{x \in X : q'(x) = 0\} \qquad (6\text{-}62)$$

l'ensemble des points de l'intervalle $[a, b]$ auxquels la dérivée s'annule;

$$S_2 = \{x \in X : q'(x) \text{ n'existe pas}\} \qquad (6\text{-}63)$$

l'ensemble des points de l'intervalle $[a, b]$ auxquels il n'y a pas de dérivée;

$$S_3 = \{a, b\} \qquad (6\text{-}64)$$

l'ensemble des points comportant les extrémités de l'intervalle $[a, b]$.

La valeur x^* qui minimise ou maximise $q(x)$ doit obligatoirement se trouver soit parmi les points de l'ensemble S_1, soit parmi les points de l'ensemble S_2, soit encore parmi les points de l'ensemble S_3. Autrement dit, la valeur de $x = x^*$ doit être recherchée parmi les points $\bar{x}$ de l'ensemble S défini par la condition

$$S = S_1 \cup S_2 \cup S_3 \qquad (6\text{-}65)$$

et représentant un ensemble fini. Cette règle ramène le problème de la recherche de la valeur optimale x^* à l'examen du cas précédent d'un ensemble fini des solutions admissibles.

Exemple 6-2. Pour la fonction représentée sur la figure 6-5, on a $S_1 = \{x_1\}$, $S_2 = \{x_2\}$, $S_3 = \{a, b\}$.
Donc, $S = \{a, x_1, x_2, b\}$.

Exemple 6-3. Soient $X = [0, 2]$, $q(x) = \left|x - \frac{1}{2}\right|$. Etant donné que, suivant la définition de la valeur absolue d'un y réel,

$$|y| = \begin{cases} -y & \text{pour } y < 0; \\ y & \text{pour } y \geqslant 0, \end{cases}$$

$q(x)$ peut se mettre sous la forme

$$q(x) = \begin{cases} \frac{1}{2} - x & \text{pour } x < \frac{1}{2}; \\ x - \frac{1}{2} & \text{pour } x \geqslant \frac{1}{2}. \end{cases}$$

En dérivant par rapport à x, on a:

$$q'(x) = \begin{cases} -1 & \text{pour } x < \frac{1}{2}; \\ +1 & \text{pour } x > \frac{1}{2}. \end{cases}$$

Au point $x = 1/2$ il n'y a pas de dérivée, car les valeurs de $q'(x)$ au voisinage de ce point sont différentes à gauche et à droite. On voit donc que sur l'intervalle $[0, 2]$ il n'y a pas de valeurs de x pour lesquelles $q'(x) = 0$. De cette façon, $S_1 = \varnothing$, $S_2 = \{1/2\}$, $S_3 = \{0, 2\}$, tandis que $S = \{0, 1/2, 2\}$.

En calculant les valeurs de $q(x)$ pour tous les $x \in S$, on obtient $\min\limits_{x \in [0,2]} q(x) = 0$ pour $x = 1/2$, $\max\limits_{x \in [0,2]} q(x) = 3/2$ pour $x = 2$.

On remarquera que les valeurs de $\bar{x} \in S$ définissent des minima ou des maxima relatifs ou locaux, c.-à-d. des points $\bar{x}$ de l'ensemble X tels que

$$q(\bar{x}) < q(x) \quad \text{ou} \quad q(\bar{x}) > q(x)$$

pour tous les x qui sont assez proches de $\bar{x}$. Il est vrai que parmi les points $\bar{x} \in S$ on peut trouver des points où il n'y a ni maximum relatif ni minimum relatif (par exemple, les points d'inflexion). Le calcul des valeurs de $q(x)$ aux points $x = \bar{x} \in S$ a pour but de trouver le minimum ou le maximum absolu de la fonction $q(x)$. Dans ce cas, la résolution exposée plus haut peut s'avérer assez compliquée si le nombre d'éléments de l'ensemble S est élevé et si le calcul de la fonction $q(x)$ n'est pas aisé.

Pour simplifier la résolution, on diminue le nombre d'éléments de l'ensemble S, ce qui s'obtient d'habitude par passage au calcul des dérivées de la fonction $q(x)$ au lieu de calculer la fonction elle-même (si l'expression des dérivées est plus simple que l'expression de la fonction). Dans le même but, on peut utiliser des procédés connus dans les mathématiques supérieures et concernant l'élimination de certains éléments de l'ensemble S par comparaison des signes des dérivées ou par calcul des dérivées d'ordre supérieur [48].

Exemple 6-4. Trouver la valeur de $x = x^*$ qui maximise la fonction $q(x) = e^{-x^2}$ sur l'intervalle $X = [-1, +1]$.

A partir de la condition $q'(x) = -2xe^{-x^2} = 0$ on trouve $S_1 = \{0\}$. Etant donné que la fonction $q(x)$ est continue, $S_2 = \varnothing$, $S_3 = \{-1, +1\}$. Ainsi, $S = \{-1, 0, +1\}$.

Essayons maintenant de diminuer le nombre d'éléments de l'ensemble S. Vu que

$$q'(x) = \begin{cases} > 0 & \text{pour} \quad x < 0; \\ < 0 & \text{pour} \quad x > 0, \end{cases}$$

on ne peut pas avoir de maximum aux points de l'ensemble S_3, tandis que les conditions du maximum sont satisfaites pour le point $x = 0$. Donc, $S' = S_1 = \{0\}$, $x^* = 0$, $q(x^*) = e^0 = 1$.

L'exemple qui suit est donné pour examiner les difficultés qui surgissent lorsque les variables du problème sont exprimées par des nombres entiers.

Exemple 6-5. Trouver la valeur de $i = i^*$ qui maximise la fonction $q(i) = 100i - 3i^2$ sur l'ensemble $X = \{i \in R : i$ est un entier, $0 \leqslant i \leqslant 100\}$.

Ici, le domaine X des valeurs admissibles de la variable représente un ensemble fini. Toutefois, le nombre d'éléments qu'il contient est si grand qu'il est difficile de comparer directement les valeurs de $q(i)$ obtenues pour tous les éléments de l'ensemble X. Dans des cas pareils, il faut toujours tâcher d'élargir le domaine des valeurs admissibles de la variable de façon qu'il représente un intervalle continu. On peut, par exemple, substituer à X l'ensemble $X_1 = \{x \in R : 0 \leqslant x \leqslant 100\}$, de sorte que $X \subset X_1$. Maintenant, si à la place de $q(i)$ on construit une nouvelle fonction $q_1(x)$ telle que, pour $x = i$, on a $q_1(x) = q(i)$, le problème de la recherche de $x^* \in X$ qui maximise $q(i)$ peut être remplacé par celui de la recherche de $x^* \in X_1$ qui maximise $q_1(x)$.

Posons dans l'exemple qui nous intéresse $q_1(x) = 100x - 3x^2$. Alors, $S_1 = \{16,7\}$, $S_2 = \varnothing$, $S_3 = \{0;\ 100\}$. De cette façon, $S = \{0;\ 16,7;\ 100\}$ et $x^* = 16,7$. Mais la valeur x^* obtenue n'est pas un élément de l'ensemble X. C'est pourquoi la valeur optimale i^* doit être recherchée parmi les éléments de l'ensemble X voisins de x^*. Ainsi, on obtient $i^* \in \{16;\ 17\}$, d'où l'on trouve $i^* = 17$, $q(i^*) = 833$.

d) Utilisation des formules d'interpolation

Dans toute une série de cas (par exemple lors de l'optimisation des étapes isolées d'un processus à étapes multiples), la fonction $q(x)$ n'est pas donnée par une formule et l'on n'indique que la procédure d'obtention des valeurs de $q(x)$. Dans des problèmes pareils, il n'est pas possible d'appliquer les méthodes analytiques décrites dans la division précédente, et pour pouvoir juger du caractère de la fonction $q(x)$, il faut trouver toutes les valeurs qu'elle prend sur l'intervalle $X = [a, b]$ des valeurs admissibles de la variable. Etant donné qu'il est impossible de dresser un tableau complet des valeurs de la fonction $q(x)$, on calcule $q(x)$ pour un certain sous-ensemble fini $\mathring{X}$ de l'ensemble X pour utiliser ensuite les résultats obtenus dans le but d'étudier le comportement de cette fonction sur tout l'intervalle $[a, b]$. Cette procédure s'appelle *interpolation* de la fonction $q(x)$. Sans nous attarder sur les questions générales de la théorie de l'interpolation [49], nous nous limiterons à examiner une des formules d'interpolation les plus répandues appelée formule de Newton.

Supposons que l'on ait pris sur l'intervalle $[a, b]$ $n + 1$ valeurs discrètes équidistantes de x constituant l'ensemble $\mathring{X} = \{x_0 = a, x_1, \ldots, x_n = b\}$, avec $x_i = x_0 + i\delta$, $i = 0, 1, \ldots, n$, où δ représente l'intervalle qui sépare deux valeurs discrètes voisines.

Supposons les valeurs de $q(x)$ pour $x \in \mathring{X}$ calculées et égales à $y_i = q(x_i)$, $i = 0, 1, \ldots, n$. L'interpolation consiste dans le choix d'un polynôme $p(x)$ de degré non supérieur à n, qui prend aux points $x = x_i$ les valeurs $p(x_i) = y_i$. Les valeurs de ce polynôme pour tout $x \in [a, b]$ sont prises en tant que valeurs de $q(x)$.

Ecrivons le polynôme $p(x)$ sous la forme

$$p(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \\ + a_3(x - x_0)(x - x_1)(x - x_2) + \ldots + a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}). \tag{6-66}$$

Définissons les coefficients $a_0, a_1, \ldots, a_n$ de façon que

$$p(x_0) = y_0, \quad p(x_1) = y_1, \ldots, p(x_n) = y_n.$$

En portant successivement dans (6-66) à la place de x les valeurs $x_0, x_1, \ldots, x_n$, on obtient:

$$y_0 = a_0 \quad \text{ou} \quad a_0 = y_0;$$

$$y_1 = a_0 + a_1(x_1 - x_0) = y_0 + a_1\delta \quad \text{ou} \quad a_1 = \frac{y_1 - y_0}{\delta};$$

$$y_2 = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) = \\ = y_0 + 2(y_1 - y_0) + 2\delta^2 a_2 \quad \text{ou} \quad a_2 = \frac{y_2 - 2y_1 + y_0}{2\delta^2}.$$

Ce processus de calcul successif des coefficients peut être poursuivi.

Pour réaliser l'écriture la plus commode des coefficients, il est rationnel d'introduire les désignations spéciales pour les différences des valeurs prises par la fonction. On appelle différences premières les quantités

$$\Delta y_k = y_{k+1} - y_k, \quad k = 0, \quad 1, \ldots, n-1.$$

Les différences des différences premières sont appelées différences secondes et se notent:

$$\Delta^2 y_k = \Delta y_{k+1} - \Delta y_k = y_{k+2} - 2y_{k+1} + y_k.$$

D'une façon analogue on introduit les différences troisièmes, quatrièmes, etc. Par exemple,

$$\Delta^3 y_k = \Delta^2 y_{k+1} - \Delta^2 y_k = y_{k+3} - 3y_{k+2} + 3y_{k+1} - y_k.$$

Compte tenu de ces désignations, les formules des coefficients prennent la forme

$$a_0 = y_0; \quad a_1 = \frac{\Delta y_0}{\delta}; \quad a_2 = \frac{\Delta^2 y_0}{2!\,\delta^2}; \ldots; \quad a_n = \frac{\Delta^n y_0}{n!\,\delta^n}.$$

Substituant ces valeurs dans la formule (6-66), on obtient la formule d'interpolation de Newton:

$$q(x) \approx p(x) = y_0 + \Delta y_0 \left(\frac{x - x_0}{\delta}\right) + \frac{\Delta^2 y_0}{2!}\left(\frac{x - x_0}{\delta}\right)\left(\frac{x - x_1}{\delta}\right) + \ldots \\ \ldots + \frac{\Delta^n y_0}{n!}\left(\frac{x - x_0}{\delta}\right)\left(\frac{x - x_1}{\delta}\right)\ldots\left(\frac{x - x_{n-1}}{\delta}\right). \tag{6-67}$$

D'habitude, cette formule s'écrit d'une façon différente, en posant

$$\frac{x-x_0}{\delta}=u \quad \text{ou} \quad x=x_0+\delta u. \tag{6-68}$$

Compte tenu de (6-68), la formule de Newton prend la forme

$$q(x)\approx p(x_0+\delta u)=y_0+u\Delta y_0+\frac{u(u-1)}{2!}\Delta^2 y_0+\ldots$$
$$\ldots+\frac{u(u-1)(u-2)\ldots(u-n+1)}{n!}\Delta^n y_0. \tag{6-69}$$

Dans certains cas, l'interpolation est commode à réaliser à l'aide de la formule

$$p(x)=\sum_{j=0}^{n} y_j \prod_{\substack{i=0\\ i\neq j}}^{n}\frac{(x-x_i)}{(x_j-x_i)}. \tag{6-70}$$

Il est évident que cette formule est vraie tout d'abord parce qu'elle représente un polynôme en x; d'autre part, pour $x=x_k$, $k=0, 1, \ldots, n$, les valeurs de $p(x_k)$ sont égales à y_k.

CHAPITRE 7

PROGRAMMATION LINÉAIRE

7-1. POSITION DU PROBLÈME DE RROGRAMMATION LINÉAIRE

a) Définitions principales

Au chapitre 6 on a déjà indiqué que le terme « programmation linéaire » se rapporte à l'étude et à la résolution du problème suivant [50 à 53].

Soit un système de m équations linéairement indépendantes à n inconnues $x_1, \ldots, x_n$ appelé *système de contraintes* du problème de programmation linéaire :

$$\left.\begin{array}{l} a_{11}x_1 + \ldots + a_{1n}x_n = b_1; \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + \ldots + a_{mn}x_n = b_m. \end{array}\right\} \qquad (7\text{-}1)$$

Le problème donné est caractérisé par le fait que le nombre d'équations est inférieur au nombre d'inconnues, c.-à-d. $m < n$. On demande de trouver les valeurs non négatives des variables ($x_i \geqslant 0$, $i = 1, \ldots, n$) qui vérifient les équations (7-1) et minimisent la fonction objectif

$$q = c_0 + c_1x_1 + \ldots + c_nx_n, \qquad (7\text{-}2)$$

appelée souvent *forme linéaire.*

Donnons quelques explications concernant ce problème. Si le nombre d'équations est égal au nombre d'inconnues ($m = n$), le système d'équations (7-1) est examiné dans l'algèbre ordinaire. Si, dans ce cas, le déterminant du système n'est pas nul, la solution est unique et on la trouve par des procédés bien étudiés. Mais si le nombre d'équations est inférieur au nombre d'inconnues ($m < n$), le système d'équations (7-1) possède une infinité de solutions et on a une infinité de collections de variables x_i ($i = 1, \ldots, n$) qui vérifient les équations (7-1). Chaque collection de variables x_i ($i = 1, \ldots, n$) satisfaisant au système d'équations (7-1) sera appelée solution.

Mais aux variables x_i on impose une contrainte supplémentaire en vertu de laquelle ces variables doivent être non négatives ($x_i \geqslant 0$). Dans le cas général, il y a une infinité de solutions satisfaisant à cette condition supplémentaire. Toute solution du système (7-1) comportant des valeurs non négatives des variables sera appelée solution admissible. Ainsi, l'essence du problème de programmation linéaire consiste à choisir dans la collection de solutions admissibles celle qui minimise la forme linéaire (7-2).

Parfois, on rencontre des problèmes de programmation linéaire dont toutes ou une partie des équations du type (7-1) se présentent sous la forme d'inégalités. Ainsi, à la place de l'équation

$$a_{j1}x_1 + \ldots + a_{jn}x_n = b_j \tag{7-3}$$

le système (7-1) peut contenir une inégalité de la forme

$$a_{j1}x_1 + \ldots + a_{jn}x_n \leqslant b_j \tag{7-4}$$

ou

$$a_{j1}x_1 + \ldots + a_{jn}x_n \geqslant b_j. \tag{7-5}$$

Toutefois, ces inégalités sont sans difficulté mises sous forme d'équations par l'introduction d'une variable additionnelle $x_{n+j} \geqslant 0$ de façon que l'une des deux expressions ci-après ait lieu en fonction du sens de l'inégalité:

$$\left.\begin{aligned} a_{j1}x_n + \ldots + a_{jn}x_n + x_{n+j} = b_j; \\ a_{j1}x_n + \ldots + a_{jn}x_n - x_{n+j} = b_j. \end{aligned}\right\} \tag{7-6}$$

Cette transformation ne fait qu'augmenter le nombre de variables, ce qui ne change pas l'essence du problème.

Parfois, on ne demande pas de minimiser la forme linéaire (7-2) mais de la maximiser. Ce problème est ramené au problème précédent en inversant le signe de l'expression de q. Vu que dans ce qui suit on rencontrera des problèmes des deux types, convenons de désigner la valeur de la fonction objectif par q s'il faut la minimiser, et par q' s'il s'agit de sa maximisation.

Avant de chercher la solution des équations (7-1), satisfaisant à toutes les exigences imposées ($x_i \geqslant 0$, $i = 1, \ldots, n$; $q = \min$), essayons de trouver une solution quelconque de ces équations. Etant donné que le nombre de variables n de ce système est supérieur au nombre d'équations m, on peut trouver une des solutions possibles en égalant à zéro $n - m$ variables quelconques. On obtient ainsi un système de m équations à m inconnues qui peut être résolu par des méthodes ordinaires connues de l'algèbre. Il est vrai que, pour que le système de m équations à m inconnues admette une solution, le déterminant composé à partir des coefficients des inconnues ne doit pas être nul. Si cette condition n'est pas satisfaite, on peut égaler à zéro autres $n - m$ variables. La solution obtenue de cette façon est appelée solution *de base*. On peut maintenant introduire certains termes largement utilisés dans les problèmes de programmation linéaire.

On appelle *base* toute collection de m variables telles que le déterminant composé de leurs coefficients ne soit nas nul. Ces m variables sont dites variables *de base* (par rapport à la base donnée). Les $n - m$ variables restantes sont des variables *libres* (ou *secondai-*

res). Chaque système d'équations (7-1) concret peut avoir plusieurs bases différentes dont chacune contient ses variables de base.

Si l'on suppose nulles toutes les variables libres et si l'on résout le système de m équations à m inconnues ainsi obtenu, on aboutit à une solution *de base*. Toutefois, parmi les différentes solutions de base on trouvera des solutions qui donnent des valeurs négatives de certaines variables. Ces solutions de base sont en contradiction avec les données du problème, donc elles sont *inadmissibles*.

Une solution de base *admissible* est une solution de base qui en même temps est admissible, c.-à-d. donne des valeurs non négatives des variables de base. Les solutions de base admissibles sont les plus simples parmi les solutions admissibles du système (7-1). Mais une condition supplémentaire est imposée à la solution du problème: la forme linéaire (7-2) doit atteindre son minimum pour la solution trouvée. Compte tenu de cette condition supplémentaire, on voit que la résolution du problème devient plus compliquée, mais, toutefois, la notion de solution de base admissible joue aussi un rôle très important lors de la recherche de la solution complète du problème.

b) Exemples de problèmes de programmation linéaire

La programmation linéaire a pris naissance lorsqu'on s'est proposé de trouver les variantes les plus avantageuses de résolution des différents problèmes liés à la planification et à la production. Ces problèmes se caractérisent par une grande liberté de variation des différents paramètres et par une série de contraintes. On demande de trouver des valeurs des paramètres telles qu'à un certain point de vue elles soient les meilleures. Ces problèmes concernent la recherche du plus rationnel mode d'utilisation des matières premières et des matériaux, la détermination des régimes de production les plus avantageux, l'augmentation de l'efficacité des moyens de transport, etc. Si la commande de la production est automatisée, ces problèmes doivent être résolus automatiquement et continûment. Pour cette raison la connaissance des méthodes de résolution de ces problèmes est nécessaire à n'importe quel ingénieur et devient particulièrement importante pour un spécialiste de l'automation.

Les exemples qui vont suivre nous permettront d'illustrer l'essence des problèmes de programmation linéaire.

Exemple 7-1. Problème d'utilisation des ressources. Pour réaliser l processus technologiques différents $T_1, \ldots, T_l$, une usine a besoin de m genres de ressources $S_1, \ldots, S_m$ (matières premières, combustible, matériaux, outillage, etc.). Les réserves de chaque genre de ces ressources sont limitées et sont égales à $b_1, \ldots, b_m$. On connaît la consommation des ressources par unité de production pour chaque processus technologique. Déterminer le volume de chaque type de production pour que le profit apporté par la réalisation de cette production soit maximal.

Désignons par a_{ij} la consommation de ressources du genre S_i par unité de production du type T_j, et par c_j le profit apporté par la réalisation d'une unité de production du type T_j. Toutes les données dont on dispose sont concentrées dans le tableau 7-1, où, pour concrétiser, l'on pose $l = 3$ et $m = 4$.

Tableau 7-1

Systématisation des données initiales du problème d'utilisation des ressources

Genre de ressources	Consommation de ressources par unité de production			Réserves de ressources
	T_1	T_2	T_3	
S_1	a_{11}	a_{12}	a_{13}	b_1
S_2	a_{21}	a_{22}	a_{23}	b_2
S_3	a_{31}	a_{32}	a_{33}	b_3
S_4	a_{41}	a_{42}	a_{43}	b_4
Profit par unité de production réalisée	c_1	c_2	c_3	

Désignons par x_j le nombre d'unités de production du type T_j fabriquées. Les contraintes de ce problème imposent que la consommation des ressources du genre S_i ($i = 1, \ldots, m$) pour la fabrication de tous les types de production ne dépasse pas les réserves disponibles:

$$\sum_{j=1}^{l} a_{ij}x_j \leqslant b_i, \quad i=1, \ldots, m. \tag{7-7}$$

Il est aisé de transformer ces contraintes en équations en introduisant les variables $x_{l+i} \geqslant 0$ qui représentent les ressources non utilisées du genre S_i. De cette façon, au lieu de (7-7) on obtient:

$$\sum_{j=1}^{l} a_{ij}x_j + x_{l+i} = b_i, \quad i=1, \ldots, m. \tag{7-8}$$

Le profit que donne la réalisation de la production sera

$$q' = \sum_{j=1}^{l} c_j x_j. \tag{7-9}$$

Le plan de fabrication optimal sera donné par la solution non négative du système d'équations (7-8) pour laquelle la fonction objectif (7-9) sera maximale.

Exemple 7-2. Problème de répartition de la fabrication entre plusieurs entreprises. Le plan de la branche prévoit la fabrication durant la période T des types de production suivants:

A_1 . . . . N_1 pièces;
A_2 . . . . N_2 pièces;
. . . .
A_l . . . . N_l pièces.

Ces types de production peuvent être fabriqués par r entreprises homogènes $E_1, \ldots, E_r$. Supposons qu'aucune entreprise ne soit en état de fabriquer plusieurs types de production à la fois. On donne en outre:

a_{ij} est la quantité de la production A_i fabriquée par l'entreprise E_j par unité de temps;

b_{ij}, le coût de l'unité de production du type A_i fabriquée par l'entreprise E_j;

x_{ij}, la période durant laquelle l'entreprise E_j fabrique de la production A_i.

Trouver les valeurs de x_{ij} assurant le coût minimal de la production.

Contraintes:

1) le temps de travail de chaque entreprise ne doit pas dépasser T

$$\sum_{i=1}^{l} x_{ij} \leqslant T, \quad j=1, \ldots, r; \tag{7-10}$$

2) la quantité de production doit correspondre à la nomenclature

$$\sum_{j=1}^{r} a_{ij}x_{ij} = N_i, \quad i=1, \ldots, l. \tag{7-11}$$

La fonction objectif représentera le coût total de la production fabriquée. Tenant compte du fait que la quantité $a_{ij}b_{ij}x_{ij}$ représente le coût de la partie de production A_i fabriquée par l'entreprise E_j, on voit que le coût total de la production fabriquée sera

$$q = \sum_{j=1}^{r} \sum_{i=1}^{l} a_{ij}b_{ij}x_{ij}. \tag{7-12}$$

Conformément aux données du problème, cette quantité doit être minimisée en respectant les contraintes (7-10) et (7-11).

Exemple 7-3. Problème de transport. Aux points $P_1, \ldots, P_l$ se trouvent les charges homogènes en quantités $a_1, \ldots, a_l$. Les quantités $b_1, \ldots, b_r$ de ces charges doivent être transportées aux points $Q_1, \ldots, Q_r$ de façon que le coût total du transport soit minimal, en supposant que les charges à transporter sont égales aux stocks dont on dispose

$$\sum_{i=1}^{l} a_i = \sum_{j=1}^{r} b_j. \tag{7-13}$$

Désignons par x_{ij} la charge transportée de P_i à Q_j et par c_{ij} le coût du transport de l'unité de cette charge. Les contraintes imposées sont les suivantes:

1) la charge expédiée du point P_i en tous les points de destination doit être égale aux stocks a_i dont on dispose:

$$\sum_{j=1}^{r} x_{ij} = a_i, \quad i=1, \ldots, l; \tag{7-14}$$

2) la charge arrivée à Q_j en provenance de tous les points d'expédition doit être égale à la quantité requise b_j:

$$\sum_{i=1}^{l} x_{ij} = b_j, \quad j=1, \ldots, r. \tag{7-15}$$

La fonction objectif définit le coût total du transport de toutes les charges

$$q=\sum_{j=1}^{r}\sum_{i=1}^{l}c_{ij}x_{ij}. \qquad (7\text{-}16)$$

Exemple 7-4. Problème du choix de l'appareillage correspondant à la variante optimale. On demande de projeter une calculatrice numérique capable d'effectuer successivement r opérations mathématiques. Elle doit donc comprendre r blocs disposés en série. Chacun de ces blocs peut être réalisé en une de l versions possibles: à tubes électroniques, à éléments semi-conducteurs, à transistors et noyaux de ferrite, à micromodules, etc. Les contraintes imposées concernent le coût maximal (X), les cotes d'encombrement maximales (Y) et le temps maximal des opérations (Z). Choisir la variante la plus avantageuse en ce qui concerne les exigences formulées.

Désignons respectivement par x_i, y_i et z_i le coût, l'encombrement et la durée de l'opération de l'i-ème bloc. Alors, les contraintes peuvent s'écrire

$$\sum_{i=1}^{r}x_i\leqslant X, \qquad \sum_{i=1}^{r}y_i\leqslant Y, \qquad \sum_{i=1}^{r}z_i\leqslant Z. \qquad (7\text{-}17)$$

Les grandeurs x_i, y_i et z_i sont fonction de la version du bloc, donc elles sont des éléments des ensembles suivants:

$$x_i\in\{x_{i1},\ldots,x_{il}\}; \quad y_i\in\{y_{i1},\ldots,y_{il}\}; \quad z_i\in\{z_{i1},\ldots,z_{il}\}, \qquad (7\text{-}18)$$

où x_{ij}, y_{ij}, z_{ij} désignent respectivement le coût, l'encombrement et la durée de l'opération de l'i-ème bloc exécuté en la j-ième version.

Si c_1, c_2 et c_3 sont des coefficients qui caractérisent la valeur relative de la diminution du coût, de l'encombrement et de la durée d'une opération, la condition qui définit la variante optimale de l'appareillage va s'écrire

$$q=c_1\sum_{i=1}^{r}x_i+c_2\sum_{i=1}^{r}y_i+c_3\sum_{i=1}^{r}z_i=\min. \qquad (7\text{-}19)$$

Le présent problème diffère du problème de programmation linéaire formulé au début de ce chapitre par le fait que les variables x_i, y_i, z_i ne peuvent pas prendre n'importe quelles valeurs non négatives mais seulement des valeurs faisant partie des ensembles finis (7-18), bien que les contraintes (7-17) et la fonction objectif (7-19) soient linéaires. C'est pourquoi les méthodes ordinaires de la programmation linéaire ne sont pas applicables au problème en question et il faut recourir aux méthodes de la programmation en nombres entiers (ou programmation discrète).

c) Interprétation géométrique du problème de programmation linéaire

Pour se faire une idée plus complète du problème de programmation linéaire, on donne ci-dessous l'interprétation géométrique d'un problème de ce type. Soient donnés le système d'équations

$$\left.\begin{aligned}-2x_1+x_2+x_3&=2;\\ x_1-2x_2+x_4&=2;\\ x_1+x_2+x_5&=5\end{aligned}\right\} \qquad (7\text{-}20)$$

et la fonction objectif

$$q = x_2 - x_1. \tag{7-21}$$

On demande de trouver les valeurs non négatives des variables vérifiant les équations (7-20) et minimisant la fonction objectif (7-21). Il s'avère plus commode de donner les règles de résolution du problème de programmation linéaire pour le cas où il faut maximiser la fonction objectif et, pour cela, prenons en qualité de fonction objectif l'expression

$$q' = -q = x_1 - x_2. \tag{7-22}$$

Dans l'exemple considéré, le nombre d'équations $m = 3$ et le nombre d'inconnues $n = 5$, de sorte que l'on a $m = 3$ variables de base et $n - m = 2$ variables libres. Le fait d'avoir seulement deux variables libres nous permet d'illustrer la solution géométrique du problème dans un espace à deux dimensions, donc dans le plan.

Le système de trois équations (7-20) peut être résolu par rapport à trois variables, par exemple, par rapport à x_3, x_4 et x_5 en les exprimant par l'intermédiaire de x_1 et x_2. Ainsi on obtient :

$$\left.\begin{aligned} x_3 &= 2 + 2x_1 - x_2; \\ x_4 &= 2 - x_1 + 2x_2; \\ x_5 &= 5 - x_1 - x_2. \end{aligned}\right\} \tag{7-23}$$

D'après les données du problème de programmation linéaire, les variables peuvent prendre seulement des valeurs non négatives, ce qui signifie que le domaine des valeurs admissibles des variables sera déterminé par les conditions

$$x_i \geqslant 0, \qquad i = 1, 2, 3, 4, 5. \tag{7-24}$$

Chacune des inégalités (7-24) définit un certain demi-plan dans le plan (x_1, x_2). Ainsi, l'inégalité $x_1 \geqslant 0$ définit le demi-plan supérieur, tandis que l'inégalité $x_3 \geqslant 0$ définit le demi-plan situé d'un côté de la droite $2 + 2x_1 - x_2 = 0$ et notamment celui qui contient l'origine des coordonnées, ce que l'on vérifie facilement en substituant dans la première des inégalités (7-23) les coordonnées du point (0, 0). Le domaine correspondant à $x_3 < 0$ est *interdit* et sur la figure 7-1 il est marqué par la hachure.

Sur la figure 7-2 sont représentées les constructions analogues exécutées pour tous les x_i. Les droites correspondant à la condition $x_i = 0$ ($i = 1, 2, 3, 4, 5$) y sont marquées du chiffre (i). Les domaines interdits correspondant à $x_i < 0$ sont marqués par la hachure. De la figure 7-2 on voit que le domaine des valeurs admissibles des variables x_1 et x_2 est représenté par le polygone *Oabcd* (figurant ici en grisé). Il importe de remarquer que le polygone des solutions

admissibles est convexe, car il représente l'intersection des domaines convexes définis par les conditions $x_i \geqslant 0$.

La construction effectuée permet de donner une interprétation géométrique de la solution de base. Etant donné que chaque droite de la figure 7-2 correspond à l'annulation de l'une des variables, aux points d'intersection de deux droites on verra s'annuler deux,

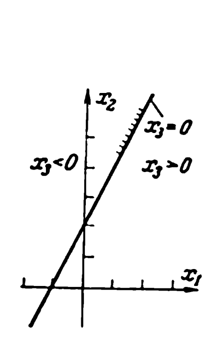


Fig. 7-1. Demi-plan $2 + 2x_1 - x_2 \geqslant 0$

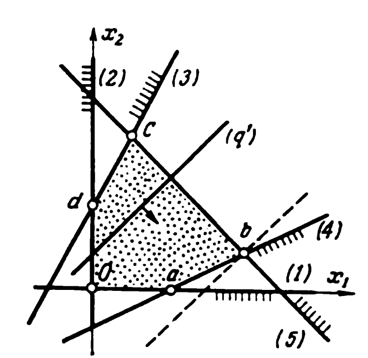


Fig. 7-2. Polygone des solutions admissibles

c.-à-d. $n - m$ variables. Mais $n - m$ représente le nombre de variables libres dont l'annulation correspond à une solution de base. Ainsi, les points d'intersection des droites $x_i = 0$ ($i = 1, \ldots, 5$) définissent les solutions de base du problème de programmation linéaire.

Parmi les solutions de base il y en a qui n'appartiennent pas au domaine des solutions admissibles. Ce sont des solutions de base inadmissibles. Seulement les points d'intersection des droites $x_i = 0$ qui sont en même temps sommets du polygone des solutions admissibles appartiennent au domaine des solutions admissibles. Il s'ensuit que les sommets du polygone des solutions admissibles correspondent aux solutions de base admissibles.

Examinons maintenant les conditions qui assurent la maximisation de la fonction objectif (7-22). Au point de vue géométrique, pour tout q', l'expression (7-22) définit une droite tracée sous un angle de 45° à l'axe des abscisses, l'augmentation de q' entraînant le déplacement de la droite dans le sens indiqué par la flèche visible sur la figure 7-2. Cette droite ne sera compatible avec la solution admissible du problème que dans le cas où elle a des points communs avec le domaine des solutions admissibles. La valeur maximale de q' sera atteinte pour la position extrême de la droite lorsque celle-ci se transforme en droite d'appui au domaine des solutions admissibles (ligne en trait interrompu sur la figure 7-2). Mais la droite d'appui à un polygone convexe passe obligatoirement par au moins un de

ses sommets qui, comme on l'a déjà vu, correspondent aux solutions de base admissibles.

On arrive ainsi à une conclusion très importante: la solution du problème de programmation linéaire qui maximise la fonction objectif q' se trouve obligatoirement parmi les solutions *de base admissibles.*

Cette conclusion est facilement généralisée au cas où m et n sont arbitraires et $n - m > 2$. Dans ce cas, les conditions $x_i \geqslant 0$ ($i = 1, \ldots, n$) définissent des demi-espaces dans un espace multidimensionnel, dont l'intersection définit le domaine des solutions admissibles sous la forme d'un polyèdre convexe. Les sommets de ce dernier correspondront aux solutions de base admissibles. L'expression de la fonction objectif définit un hyperplan dans l'espace multidimensionnel considéré qui, pour $q' = \max$, sera un hyperplan d'appui au polyèdre des solutions admissibles, ce qui signifie qu'il passera nécessairement par au moins un de ses sommets. De cette façon, dans ce cas aussi, la solution du problème de programmation linéaire se trouvera parmi les solutions *de base admissibles.*

La conclusion à laquelle on est arrivé permet de tracer la voie menant à la solution du problème de programmation linéaire. Etant donné que la solution doit se trouver parmi les solutions de base admissibles dont le nombre est fini, on peut trouver toutes les solutions de base admissibles et calculer pour chacune d'entre elles la valeur de q'. Finalement, la solution recherchée sera celle pour laquelle la valeur de q' sera maximale.

Bien que possible, cette voie de résolution du problème est difficile, car le nombre de solutions de base admissibles peut être considérable. Toutefois, il y a des méthodes rationnelles de triage successif des solutions de base qui permettent de ne pas considérer toutes les solutions de base admissibles mais seulement un nombre minimal d'entre elles. Une des méthodes de triage les plus répandues est la méthode dite du simplexe qui sera examinée dans le paragraphe qui suit.

7-2. RÉSOLUTION DU PROBLÈME DE PROGRAMMATION LINÉAIRE

a) Algèbre de la méthode du simplexe

L'essentiel de la méthode du simplexe consiste en ce qui suit. On trouve tout d'abord une solution de base admissible quelconque. Cette solution peut être trouvée en supposant libres et égales à zéro $n - m$ variables et en résolvant le système d'équations (7-1) ainsi obtenu. Si certaines des variables de base s'avèrent négatives, il faut choisir d'autres variables libres, c.-à-d. passer à une nouvelle base.

Après avoir trouvé cette solution de base admissible, il faut voir si la fonction objectif q' n'a pas encore atteint son maximum. Sinon, on recherche une nouvelle solution de base admissible en allant dans le sens d'augmentation de la valeur de la fonction objectif q'. Ensuite, la procédure est réitérée. Etant donné que la nouvelle solution de base admissible est choisie de façon à augmenter la valeur de la fonction objectif, la méthode en question permet de considérer un nombre minimal de solutions de base admissibles et aboutit rapidement à la solution finale. Analysons cette méthode plus en détail en l'appliquant à l'exemple précédemment cité.

Considérons le système d'équations (7-20) pour lequel il faut trouver une solution non négative qui maximise la fonction objectif (7-22). Vu que $n - m = 2$, on peut prendre en tant que variables libres n'importe quel couple de variables, par exemple, x_1 et x_2. En égalant à zéro ces deux variables, on a, à partir de (7-20), la solution de base $x_1 = x_2 = 0$, $x_3 = 2$, $x_4 = 2$, $x_5 = 5$ qui est admissible et qui donne $q' = x_1 - x_2 = 0$.

Il nous intéresse maintenant de voir si cette solution rend maximale la fonction objectif. A cette fin, on peut chercher une nouvelle solution de base pour laquelle la fonction objectif ait une valeur plus grande. Pour passer à une nouvelle solution de base admissible, une des variables libres (x_1 ou x_2) doit être prise en tant que variable de base. Dans ce cas, elle sera différente de zéro, ce qui signifie qu'elle augmentera. Il s'ensuit que si une quelconque des variables libres figure avec le signe « + » dans l'expression de la fonction objectif (ce qui veut dire que la fonction objectif augmente avec l'augmentation de cette variable), le maximum de la fonction objectif n'est pas atteint et la variable en question doit être prise comme variable de base différente de zéro.

Mais lorsque la variable libre s'accroît, certaines variables de base commenceront à diminuer. Les valeurs négatives des variables étant inadmissibles, il faut prendre en qualité de nouvelle variable libre celle des variables de base qui s'annule avant toutes les autres.

Dans l'exemple considéré, l'expression de la fonction objectif (7-22) contient la variable libre x_1 avec le signe « + ». Donc, le maximum de la fonction objectif n'est pas atteint et la variable x_1 doit être transformée en variable de base. Pour déterminer la nouvelle variable libre, exprimons les variables de base x_3, x_4, x_5 en fonction des variables libres en ramenant les équations (7-20) à la forme (7-23). Ces équations montrent que, pour $x_2 = 0$, l'augmentation de x_1 n'entraîne pas la diminution de x_3 mais fait diminuer x_4 et x_5, de façon que, pour $x_1 = 2$, on a $x_4 = 0$, $x_5 = 3 > 0$. Donc, la variable x_4 doit être prise comme variable libre, ce qui nous amène à la nouvelle base x_1, x_3, x_5.

Pour continuer la résolution du problème, il faut résoudre le système (7-20) par rapport aux nouvelles variables de base en le

mettant sous la forme

$$\left.\begin{aligned} x_1 &= 2 + 2x_2 - x_4\,; \\ x_3 &= 6 + 3x_2 - 2x_4\,; \\ x_5 &= 3 - 3x_2 + x_4. \end{aligned}\right\} \qquad (7\text{-}25)$$

Exprimons également la fonction objectif en fonction des nouvelles variables libres x_2 et x_4:

$$q' = 2 + x_2 - x_4. \qquad (7\text{-}26)$$

Par des raisonnements analogues on arrive à la conclusion que le maximum de la fonction objectif n'est pas atteint et qu'il faut transformer la variable libre x_2 en variable de base et la variable de base x_5 en variable libre. Dans ce cas, la nouvelle base comportera les variables x_1, x_2, x_3.

En résolvant les équations (7-20) par rapport aux nouvelles variables de base, on obtient:

$$\left.\begin{aligned} x_1 &= 4 - \frac{1}{3}x_4 - \frac{2}{3}x_5\,; \\ x_2 &= 1 + \frac{1}{3}x_4 - \frac{1}{3}x_5\,; \\ x_3 &= 9 - x_4 - x_5, \end{aligned}\right\} \qquad (7\text{-}27)$$

et la fonction objectif prend la forme:

$$q' = 3 - \frac{2}{3}x_4 - \frac{1}{3}x_5. \qquad (7\text{-}28)$$

L'expression (7-28) montre que l'augmentation des variables libres x_4 et x_5 entraîne la diminution de la valeur de q'. Il s'ensuit que, pour la base donnée, la fonction objectif q' atteint son maximum et la solution du problème considéré est fournie par la collection de variables suivante:

$$x_4 = x_5 = 0, \quad x_1 = 4, \quad x_2 = 1, \quad x_3 = 9.$$

Dans ce cas, $q' = -q = 3$, $q = -3$.

La méthode de résolution examinée implique des transformations encombrantes du système d'équations linéaires, ce qui constitue son principal inconvénient. Ces transformations peuvent être simplifiées pour beaucoup en mettant les équations sous la forme des tableaux contenant les coefficients des variables. Cela permet de passer d'un système d'équations à un autre en recalculant tout simplement les coefficients portés aux tableaux suivant des règles purement formelles qui, d'autre part, se prêtent bien à l'utilisation des calculatrices électroniques.

b) Méthode du tableau du simplexe : recherche de la solution optimale

Lors de la mise en œuvre de cette méthode, il est commode d'utiliser une forme spéciale d'écriture des équations (7-1) et de la fonction objectif (7-2). Désignons par x'_i, $i = 1, \ldots, m$, les variables de base et par x''_j, $j = 1, \ldots, (n - m)$ les variables libres. En exprimant la fonction objectif et les variables de base en fonction des variables libres, formulons le problème de programmation linéaire de la façon suivante:

Il faut maximiser

$$q' = -q = \alpha_{00} - \sum_{j=1}^{n-m} \alpha_{0j} x''_j \tag{7-29}$$

sous les contraintes

$$x'_i = \alpha_{i0} - \sum_{j=1}^{n-m} \alpha_{ij} x''_j ; \quad i = 1, \ldots, m ; \quad x'_i \geqslant 0 ; \quad x''_j \geqslant 0. \tag{7-30}$$

Cette écriture permet de présenter le problème sous la forme de la matrice suivante des coefficients des variables:

$$\begin{array}{c} \\ q' \\ x'_1 \\ \cdots \\ x'_m \end{array} \begin{array}{c} 1 \; -x''_1 \; \ldots \; -x''_{n-m} \\ \left\| \begin{array}{ccc} \alpha_{00}\alpha_{01} & \ldots & \alpha_{0(n-m)} \\ \alpha_{10}\alpha_{11} & \ldots & \alpha_{1(n-m)} \\ \ldots & \ldots & \ldots \\ \alpha_{m0}\alpha_{m1} & \ldots & \alpha_{m(n-m)} \end{array} \right\| \end{array} . \tag{7-31}$$

La forme des coefficients de la matrice (7-31) permet d'apprécier sans difficulté si la solution de base trouvée est admissible, et, dans l'affirmative, si elle est optimale. En effet, on remarque que la colonne des coefficients α_{i0}, $i \neq 0$, représente la solution de base correspondant à la base $x'_1, \ldots, x'_m$, et que la ligne des coefficients α_{0j}, $j \neq 0$, est composée des coefficients changés de signe des variables libres figurant dans l'expression de q', ce qui permet de tirer la conclusion que la solution de base correspondant à la base $x'_1, \ldots x'_m$ est admissible si $\alpha_{i0} \geqslant 0$, $i \neq 0$. Si en outre $\alpha_{0j} \geqslant 0$, $j \neq 0$, cette solution de base est en même temps la solution optimale. Il est aussi évident que, pour la solution de base optimale, le coefficient α_{00} donne la valeur de $q' = q'_{\max} = -q_{\min}$.

Commençons la résolution du problème en trouvant une solution de base admissible quelconque qui, dans le cas général, n'est pas optimale. Présentons cette solution de base sous la forme d'un tableau de coefficients analogue à la matrice (7-31). Pour passer à une meilleure solution de base, il faut remanier la matrice des coefficients. Formulons le procédé permettant d'effectuer cette transfor-

mation sous la forme d'un système de règles que nous allons illustrer à l'aide du problème résolu au paragraphe précédent. Sans donner une argumentation rigoureuse à ces règles, on remarquera qu'elles correspondent strictement aux transformations du système d'équations effectuées dans l'exemple précédent et qui sont faciles à vérifier par confrontation des transformations correspondantes.

Supposons que le problème de programmation linéaire soit donné sous la forme du système d'équations (7-20) et de la fonction objectif (7-22) qu'il faut maximiser. En prenant x_1 et x_2 comme variables libres, ramenons ce système d'équations et la fonction objectif à la forme (7-29) et (7-30)

$$\left.\begin{aligned} q' &= 0-(-x_1+x_2);\\ x_3 &= 2-(-2x_1+x_2);\\ x_4 &= 2-(x_1-2x_2);\\ x_5 &= 5-(x_1+x_2). \end{aligned}\right\} \qquad (7\text{-}32)$$

La matrice des coefficients est donnée sous la forme du tableau 7-2, *a* dont les cases sont assez grandes pour permettre de porter dans leur coin de gauche supérieur les coefficients α_{ij} des équations (7-32).

Tableau 7-2

Transformations successives des tableaux des coefficients lors de la résolution du problème de programmation linéaire

	1	$-x_1$	$-x_2$
q'	0 2	−1 1	1 −2
x_3	2 4	−2 2	1 −4
x_4	2 2	[1] 1	−2 −2
x_5	5 −2	1 −1	1 2

a)

	1	$-x_4$	$-x_2$
q'	2 1	1 $-\frac{1}{3}$	−1 $\frac{1}{3}$
x_3	6 3	2 −1	−3 1
x_1	2 2	1 $-\frac{2}{3}$	−2 $\frac{2}{3}$
x_5	3 1	−1 $-\frac{1}{3}$	[3] $\frac{1}{3}$

b)

	1	$-x_4$	$-x_5$
q'	3	$\frac{2}{3}$	$\frac{1}{3}$
x_3	9	1	1
x_1	4	$\frac{1}{3}$	$\frac{2}{3}$
x_2	1	$-\frac{1}{3}$	$\frac{1}{3}$

c)

Voyons si l'on n'a pas encore trouvé la solution optimale vérifiant la condition $\alpha_{0j} \geqslant 0$, $j \neq 0$. Etant donné que α_{01} (le coefficient de $-x_1$ dans l'expression de q') est négatif, on constate que la solution optimale n'est pas trouvée et la variable x_1 doit être transformée en variable de base. Mettons en évidence la colonne de la variable x_1 à l'aide des barres verticales doubles. Si les coefficients α_{0j} de plusieurs variables libres s'avèrent négatifs, n'importe quelle de ces variables peut être transformée en variable de base.

Déterminons maintenant la variable de base qui doit être transformée en variable libre. Ce sera évidemment la variable qui s'annulera plus tôt avec l'accroissement de x_1, donc la variable de base x_i pour laquelle le coefficient situé dans la colonne mise en évidence $\alpha_{i1} > 0$ et le rapport α_{i0}/α_{i1} est minimal. La variable de base x_4 avec $\alpha_{40} = 2$ et $\alpha_{41} = 1$ vérifie ces conditions. La ligne correspondant à la variable de base x_4 est également marquée par des barres horizontales doubles.

Appelons *coefficient général* (*pivot*) le coefficient $\lambda = \alpha_{41}$ se trouvant au coin de gauche supérieur de la case à l'intersection de la ligne et de la colonne marquées. Dans le cas traité, $\lambda = 1$ (valeur encadrée sur le tableau).

Maintenant, il faut remplir les coins inférieurs des cases en suivant les règles ci-dessous:

1) dans la case se trouvant à l'intersection de la ligne et de la colonne marquées, on écrit $1/\lambda$;

2) dans les cases de la ligne marquée, on écrit les coefficients supérieurs multipliés par λ (les coefficients supérieurs, excepté le pivot, sont écrits en caractères gras);

3) dans les cases de la colonne marquée, on écrit les coefficients supérieurs multipliés par $-\lambda$ (les coefficients inférieurs, à l'exception de la case contenant le pivot, sont donnés en caractères gras);

4) dans les autres cases, on écrit le produit des coefficients écrits en caractères gras et figurant dans la colonne et dans la ligne à l'intersection desquelles se trouve la case donnée.

Ensuite, on commence à dresser le tableau 7-2, *b* qui diffère du tableau 7-2, *a* par le fait que la variable libre marquée x_1 est devenue variable de base, tandis que la variable de base marquée x_4 est devenue variable libre. Les coins de gauche supérieurs des cases du tableau 7-2, *b* sont remplis en respectant les règles suivantes:

1) dans la ligne et la colonne correspondant aux nouvelles variables libre et de base, on porte les coefficients inférieurs de la ligne et de la colonne marquées du tableau 7-2, *a*;

2) dans les autres cases, on écrit les sommes des coefficients se trouvant dans les cases correspondantes du tableau 7-2, *a*.

Le tableau 7-2, *b* dressé de cette façon correspond à la matrice des coefficients (7-31), la nouvelle base étant x_1, x_3, x_5. Ensuite, la procédure est réitérée.

Vu que dans le tableau 7-2, *b* le coefficient $\alpha_{02} < 0$, la solution optimale n'est pas trouvée, ce qui signifie qu'il faut remplir les coins de gauche inférieurs de ce tableau suivant les règles déjà données et passer au nouveau tableau 7-2, *c* correspondant à la base x_1, x_2, x_3. Dans ce dernier, les coefficients α_{0j}, $j \neq 0$, sont positifs et on obtient la solution optimale du problème suivant la colonne des termes constants:

$$x_1 = 4\,; \quad x_2 = 1\,; \quad x_3 = 9\,; \quad x_4 = x_5 = 0\,; \quad q'_{\max} = -q_{\min} = 3\,; \quad q_{\min} = -3.$$

c) Problème dual de programmation linéaire

Revenons au problème de la distribution des ressources examiné à l'exemple 7-1 et demandons-nous quelle est, au point de vue de l'entreprise, la valeur des ressources dont celle-ci dispose. En résolvant cette question, il faut avoir en vue que les ressources qui ne peuvent pas être utilisées à fond n'ont pour l'entreprise que peu de valeur et pour en augmenter les réserves elle ne sera pas disposée d'engager des frais ne serait-ce que peu importants. Ainsi, un équipement coûteux non utilisé dans le processus technologique présente peu de valeur pour l'entreprise. Les ressources de la majeure valeur seront celles qui limitent le plus la production, donc les profits de l'entreprise. Pour augmenter les réserves de ces ressources l'entreprise est prête à assumer des frais considérables.

C'est pourquoi on peut considérer que chaque type de ressources possède un « prix fictif » [10] qui détermine la valeur présentée pour l'entreprise par le type de ressources donné au point de vue des profits apportés par la réalisation de la production. Ce prix fictif dépend des réserves disponibles de ces ressources et de leur utilité pour la production.

Si, pour une raison ou une autre, l'entreprise borne son activité à un seul processus technologique nécessitant une consommation importante d'un certain type de ressources dont les réserves sont limitées, le prix fictif de ce type de ressources sera élevé. Mais les prix fictifs établis conformément à ce processus technologique ne seront pas les meilleurs, car en mettant en œuvre d'autres processus technologiques, on arrive à une utilisation plus rationnelle de toutes les réserves. Il s'ensuit qu'il y a des prix fictifs optimaux qui correspondent au profit maximal de l'entreprise, c.-à-d. à la distribution optimale des ressources. On voit donc que la détermination des prix fictifs optimaux est étroitement liée au problème de distribution optimale des ressources, c.-à-d. au problème de programmation linéaire décrit par le système d'équations (7-8) et par la fonction objectif (7-9). Toutefois, pour déterminer les prix fictifs optimaux, on peut élaborer un problème de programmation linéaire à part.

Désignons par u_i le prix fictif de l'unité de ressources S_i. Les valeurs de u_i doivent être telles que le prix fictif des ressources utilisées dans n'importe quel processus technologique ne soit pas inférieur au profit réalisé. En faisant appel aux désignations de l'exemple 7-1 et aux données du tableau 7-1, écrivons cette condition comme suit

$$\sum_{i=1}^{m} a_{ij}u_i \geqslant c_j, \quad j = 1, \ldots, l. \tag{7-33}$$

Si l'on introduit les variables $u_{m+j} \geqslant 0$, qui représentent le dépassement qu'a le prix fictif de l'unité de production par rapport aux profits apportés par sa réalisation, le système d'inégalités (7-33) se transforme en système d'équations suivant:

$$\sum_{i=1}^{m} a_{ij}u_i - u_{m+j} = c_j, \quad j = 1, \ldots, l. \tag{7-34}$$

Les prix fictifs optimaux seront ceux qui minimisent le prix total des ressources, c.-à-d. la grandeur

$$q^* = \sum_{i=1}^{m} b_i u_i. \tag{7-35}$$

Le système de contraintes (7-33) et la fonction objectif (7-35) représentent un nouveau problème de programmation linéaire appelé problème *dual* par rapport au problème donné à l'exemple 7-1 qui est dit problème *primal* (on dit encore problème *initial* ou *fondamental*) de programmation linéaire.

Il est aisé de remarquer que les problèmes primal et dual sont étroitement liés l'un à l'autre de la façon suivante:

si le problème primal est un problème de maximisation de la fonction objectif, le problème dual sera un problème de minimisation;

les coefficients de la fonction objectif du problème primal apparaissent comme les constantes des contraintes du problème dual;

les constantes des contraintes du problème primal figurent au problème dual comme coefficients de la fonction objectif;

les coefficients des variables des contraintes du problème dual représentent les colonnes du tableau de coefficients du problème primal;

le sens des inégalités représentant les contraintes est inversé, sauf en ce qui concerne la contrainte de non-négativité des variables.

La résolution du problème primal est étroitement liée à celle du problème dual de programmation linéaire. Pour établir cette liaison, écrivons les équations des problèmes primal et dual d'une façon différente.

Supposons libres les variables $x_1, \ldots, x_l$ du problème primal et formulons ce dernier comme suit:
maximiser

$$q' = 0 - \sum_{j=1}^{l} (-c_j)\, x_j \tag{7-36}$$

sous les contraintes

$$x_{l+i} = b_i - \sum_{j=1}^{l} a_{ij} x_j, \quad i = 1, \ldots, m. \tag{7-37}$$

A ce problème est associée une matrice de la forme

$$\begin{array}{r} \\ q' \\ x_{l+1} \\ \cdots \\ x_{l+m} \end{array} \begin{array}{c} 1 \quad -x_1 \;\ldots\; -x_l \\ \left\| \begin{array}{ccc} 0 & -c_1 & \ldots -c_l \\ b_1 & a_{11} & \ldots\; a_{1l} \\ \cdot & \cdots & \cdots \\ b_m & a_{m1} & \ldots\; a_{ml} \end{array} \right\| \end{array}. \tag{7-38}$$

Prenons $u_1, \ldots, u_m$ en qualité de variables libres du problème dual et formulons-le de la façon suivante:
minimiser

$$q^* = 0 + \sum_{i=1}^{m} b_i u_i \tag{7-39}$$

sous les contraintes

$$u_{m+j} = -c_j + \sum_{i=1}^{m} a_{ij} u_i, \quad j = 1, \ldots, l. \tag{7-40}$$

A ce problème est associée une matrice de la forme

$$\begin{array}{r} \\ q^* \\ u_{m+1} \\ \cdots \\ u_{m+l} \end{array} \begin{array}{c} 1 \quad u_1 \;\ldots\; u_m \\ \left\| \begin{array}{ccc} 0 & b_{11} & \ldots\; b_m \\ -c_1 & a_{11} & \ldots\; a_{m1} \\ \cdot & \cdots & \cdots \\ -c_l & a_{1l} & \ldots\; a_{ml} \end{array} \right\| \end{array}. \tag{7-41}$$

On voit que les colonnes de la matrice (7-41) correspondent aux lignes de la matrice (7-38). Il s'ensuit que les problèmes primal et dual peuvent être décrits par la même matrice (7-38) en lui associant la correspondance suivante entre les variables du primal et du dual:

$$\left.\begin{array}{ll} x_j \longleftrightarrow -u_{m+j}, & j = 1, \ldots, l; \\ x_{l+i} \longleftrightarrow u_i, & i = 1, \ldots, m. \end{array}\right\} \tag{7-42}$$

Il est à remarquer que toute transformation de la matrice (7-38) réalisée d'après les règles données au paragraphe précédent aboutit à une nouvelle matrice qui va décrire aussi bien le problème primal que celui dual. Il en découle que la matrice de la forme la plus générale (7-31) peut servir à la description des problèmes primal et dual de programmation linéaire. Si, dans ce cas, les éléments de la première colonne (sauf, peut-être, α_{00}) sont positifs, la matrice correspond à une solution de base admissible du problème primal. Si ce sont les éléments de la première ligne (sauf, peut-être, α_{00}) qui sont positifs, la matrice correspond à une solution admissible du problème dual. Mais si les éléments aussi bien de la première colonne que de la ligne supérieure de la matrice (7-31) sont positifs (à l'exception, peut-être, de α_{00}), alors la matrice correspond à la solution optimale de tous les deux problèmes primal et dual. Dans ce cas, le coefficient α_{00} donne la valeur de la fonction objectif qui, pour la solution optimale, est la même pour le problème primal et pour celui dual:

$$q'_{max} = q^*_{min}. \tag{7-43}$$

Exemple 7-5. Le problème de distribution des ressources est donné par le tableau 7-3. Les lignes et les colonnes supplémentaires de ce tableau contiennent la solution du problème primal et du problème dual de programmation linéaire, c.-à-d. on y trouve le plan optimal de distribution des ressources, les ressources restantes correspondant au plan optimal et les prix fictifs. Le tableau montre que le prix des ressources S_1, qui abondent, est nul. Le prix maximal caractérise les ressources S_2 qui sont utilisées dans tous les processus technologiques et dont les réserves sont limitées.

Tableau 7-3

Données initiales du problème d'utilisation des ressources

Types de ressources	Consommation de ressources par unité de production			Réserves de ressources	Ressources restantes correspondant au plan optimal	Prix fictifs
	T_1	T_2	T_3			
S_1	2	1	1	25	5,5	0
S_2	1	1	1	14	0	3,0
S_3	0	4	2	19	0	0,5
S_4	3	0	1	24	0	1,0
Profits apportés par la réalisation d'une unité de production	6	5	5			
Plan optimal	5,5	1,0	7,5			

Remarquons pour finir que les prix fictifs peuvent, dans certains cas, jouer un rôle important en qualité d'instrument de gestion. Ainsi, lorsqu'il s'agit d'une grosse entreprise ou d'une branche industrielle, où beaucoup de décisions sont

prises d'une façon indépendante par les sections et les groupements de production, il est parfois difficile d'informer chaque section sur les décisions prises par les autres sections. Dans ce cas, les prix fictifs constituent pour chaque section un bon repère qui leur permet de prendre des décisions proches de l'optimum au point de vue des intérêts de toute l'entreprise.

d) Notion de programmation en nombres entiers

Dans beaucoup de problèmes de programmation linéaire, les variables représentent des unités indivisibles. Ainsi, une entreprise ne peut pas fabriquer 7,5 avions, 4,8 turbines, etc. En présence d'un tel problème, on ajoute aux contraintes (7-1) et (7-2) une nouvelle exigeant que les variables x_i soient exprimées par des nombres entiers ou qu'elles soient des éléments d'un ensemble fini (voir l'exemple 7-4). Si l'on se réfère à l'interprétation géométrique, cela signifie que les solutions admissibles ne pourront pas être cherchées dans tout le domaine défini par les contraintes (7-1), mais seulement en des points discrets isolés de ce domaine, comme on le voit sur la figure 7-3.

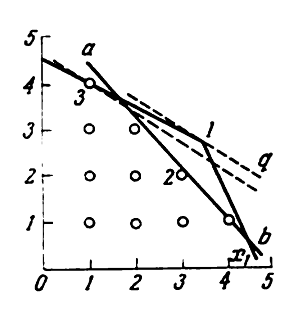


Fig. 7-3. Ensemble des solutions admissibles dans un problème de programmation en nombres entiers

On peut, bien sûr, essayer de résoudre le problème de ce type, sans tenir compte de la condition imposant une solution exprimée par un nombre entier, et trouver la solution définie par le point *1* sur la figure 7-3, pour l'arrondir ensuite à l'entier le plus proche, ce qui donnera le point *2* en tant que solution entière. Mais, en procédant de cette façon, on peut arriver à une solution qui est loin d'être optimale. Par exemple, la solution optimale en nombres entiers sur la figure 7-3 est donnée par le point *3*.

Il y a plusieurs méthodes de résolution du problème de programmation en nombres entiers parmi lesquelles la plus répandue est la méthode de Gomory. Sans nous attarder sur les calculs liés à cette méthode [45 et 46], nous n'indiquerons ci-dessous que son idée générale.

La méthode de Gomory est basée sur la méthode du simplexe à l'aide de laquelle on recherche la solution optimale sans égard pour la condition imposant que les variables soient entières ou discrètes. Si cette solution n'est pas en nombres entiers, on introduit une contrainte supplémentaire représentée sur la figure 7-3 par la ligne *ab* qui sépare une partie du domaine des solutions admissibles contenant la solution optimale trouvée et ne comportant aucun point admissible entier. Compte tenu de cette contrainte supplémentaire, la solution optimale trouvée ne sera pas comprise dans le nouveau

domaine des solutions admissibles, c.-à-d. elle sera inadmissible. C'est pourquoi le problème est à nouveau résolu par la méthode du simplexe pour trouver la solution optimale correspondant au nouveau domaine des solutions admissibles. Si cette dernière solution n'est pas entière non plus, la procédure est réitérée.

Ces derniers temps, pour la résolution du problème de programmation en nombres entiers, on applique avec succès des méthodes combinatoires dont la plus importante est celle des branches et des frontières [46].

PROBLÈMES AU CHAPITRE 7

7-1. Transformer les inégalités (7-10) en équations. Quel est le sens physique des variables qu'il faut ajouter dans ce cas?

7-2. Trouver par la méthode du simplexe le plan optimal de distribution des ressources suivant les données de l'exemple 7-5, en posant, à l'étape initiale, x_1, x_2 et x_3 comme variables libres.

7-3. Résoudre par la méthode du simplexe le problème dual correspondant aux données de l'exemple 7-5 en posant, à l'étape initiale, u_1, u_2, u_5 et u_7 comme variables libres. Comparer la matrice associée à la solution optimale avec la matrice correspondante du problème 7-2.

CHAPITRE 8

THÉORIE DES JEUX

8-1. OBJET DE LA THÉORIE DES JEUX

a) Le jeu en tant que modèle d'une situation de conflit

La théorie des jeux est une discipline mathématique à développement impétueux qui étudie les méthodes concernant la prise de décisions dans les situations dites de conflit [54, 55]. La situation est dite *de conflit* lorsque les intérêts opposés de plusieurs (d'habitude de deux) personnes s'y heurtent. Chacun des participants peut réaliser toute une série de mesures pour atteindre son but, le succès de l'un d'eux signifiant l'insuccès de l'autre.

J. von Neumann et O. Morgenstern [56], auteurs du premier traité fondamental sur la théorie des jeux, ont entrepris l'analyse des situations de conflit dans le domaine économique où, en présence d'un régime de libre concurrence, on a en tant que participants à intérêts opposés des maisons de commerce, des entreprises industrielles, etc. Pourtant, les situations de conflit se rencontrent dans beaucoup d'autres domaines. Aux situations de conflit se rapportent presque toutes les situations qui prennent naissance lors de la planification des opérations militaires, du choix d'un système d'armement, de la défense des objectifs contre les attaques, de la poursuite et de l'interception du but, et ainsi de suite. Comme exemples intéressants de situations de conflit on peut citer les compétitions sportives, les litiges réglés par arbitrage, les ventes aux enchères et les élections parlementaires où plusieurs personnes posent leurs candidatures pour un seul siège.

On voit donc que dans la pratique on a affaire à une grande variété de situations de conflit. D'habitude, il est difficile d'entreprendre une analyse directe de ces situations à cause d'un grand nombre de facteurs secondaires accessoires. Pour rendre accessible l'analyse mathématique d'une situation de conflit, cette dernière doit être simplifiée et l'on ne doit prendre en considération que les facteurs essentiels. Le modèle formalisé simplifié d'une situation de conflit est appelé *jeu*, les participants au jeu étant nommés *joueurs*. Dans ce qui suit, nous nous bornerons à l'examen des jeux à deux participants à intérêts opposés. Avant de donner une définition formelle du jeu, il est nécessaire de fournir quelques explications sur la terminologie utilisée qui pour la plupart est la même que celle utilisée dans les jeux de société (échecs, jeu de dames, jeux de cartes, etc.).

Il faut distinguer la notion de jeu de la notion de partie individuelle de ce jeu. Le jeu représente une collection de règles qui définissent le comportement des joueurs. Chaque fois où l'on fait le jeu d'une façon concrète dès son commencement jusqu'à sa fin, on dit que l'on a fait une *partie*. Les *coups* sont les éléments constitutifs du jeu. Les règles du jeu définissent la succession des coups et indiquent le caractère de chaque coup.

Il y a des coups personnels et des coups aléatoires. Le *coup personnel* représente le choix fait par le joueur d'une variante dans l'ensemble des variantes donné. Par exemple, chaque coup aux échecs est un coup personnel, le premier coup pouvant être choisi parmi 20 variantes possibles. La décision prise par le joueur qui dispose d'un coup personnel est appelée *choix*.

Le *coup aléatoire* représente également un choix d'une variante parmi un ensemble des variantes, mais, cette fois-ci, la variante n'est pas choisie par le joueur mais par un mécanisme aléatoire. Comme exemples de coups aléatoires on peut citer la donne des cartes ou le jet d'une pièce de monnaie (au jeu de pile ou face). Le choix effectué lors d'un coup aléatoire est appelé *issue* de ce coup.

Par rapport aux coups, la structure des règles est la suivante.

Les règles indiquent si le premier coup doit être un coup personnel ou un coup aléatoire. S'il s'agit d'un coup personnel, les règles énumèrent les variantes possibles et indiquent le joueur qui doit faire le choix. Si le coup est aléatoire, les règles énumèrent les variantes possibles et stipulent les probabilités de leur choix.

Pour chaque coup suivant, en fonction des choix et des issues des coups précédents, les règles déterminent:

si ce coup doit être un coup personnel ou un coup aléatoire;

les variantes possibles et les probabilités de leur choix si le coup est un coup aléatoire;

le joueur qui fait le choix, les variantes possibles parmi lesquelles on fait le choix et l'information concernant les choix et les issues des coups précédents si le coup est un coup personnel.

Et enfin, en fonction des choix et des issues des coups qui se suivent (c.-à-d. en fonction de la marche du jeu), les règles indiquent quand le jeu doit prendre fin, de même que le gain ou la perte de chacun des joueurs.

b) Notion de stratégie

Supposons que nous voulons faire une partie d'échecs avec les Blancs, mais que nous ne pouvons pas prendre personnellement part au jeu. Nous avons un suppléant qui doit faire la partie en exécutant toutes nos indications, mais qui ne sait pas jouer aux échecs et ne peut prendre indépendamment aucune décision. Pour que ce suppléant puisse faire toute la partie jusqu'à la fin, il doit

recevoir des instructions prévoyant toutes les positions possibles sur l'échiquier et déterminant le coup qui doit être joué pour chacune de ces positions. Le système complet d'instructions pareilles représente la *stratégie*.

Ainsi, la stratégie des Blancs doit indiquer le premier coup, ensuite, pour chaque réponse possible des Noirs, le coup suivant des Blancs, etc. Evidemment, la composition d'une stratégie complète au jeu d'échecs est une tâche immense pratiquement irréalisable. Par exemple, si le joueur qui joue avec les Blancs participe personnellement au jeu, il doit prendre deux décisions pour jouer les premiers deux coups, tandis que s'il fait appel à un suppléant, il doit préparer 21 décisions pour les mêmes deux coups (une décision pour le premier coup et 20 décisions pour les 20 premiers coups possibles des Noirs). Toutefois, la notion de stratégie est très utile dans beaucoup de problèmes plus simples.

De cette façon, la *stratégie* du joueur représente une description univoque de son choix dans chaque situation possible où il doit jouer un coup personnel.

Si le jeu ne comporte que des coups personnels, l'issue du jeu est définie si chacun des joueurs a choisi sa stratégie. Mais si le jeu comporte des coups aléatoires, il aura un caractère probabiliste et le choix des stratégies des joueurs ne pourra pas déterminer définitivement l'issue du jeu.

c) Description formelle d'un jeu à deux personnes

Désignons par X et Y les ensembles ou les espaces contenant toutes les stratégies possibles mises à la disposition des participants au jeu qui, dans ce qui suit, seront appelés respectivement premier et deuxième joueurs. Les grandeurs $x \in X$ et $y \in Y$ représenteront les stratégies concrètes du premier et du deuxième joueur.

Pour examiner les coups aléatoires, il est commode de considérer qu'un troisième joueur participe au jeu et qu'il utilise un mécanisme aléatoire pour jouer ses coups aléatoires. Désignons par H l'espace des stratégies de ce joueur. Toute stratégie $h \in H$ du troisième joueur constituée par la suite concrète de tous les coups aléatoires de la partie se rencontrera avec une certaine probabilité $p(h)$ facile à calculer lorsqu'on connaît la probabilité de chaque coup aléatoire dans cette suite. On voit facilement que $p(h)$ représente la distribution de probabilités sur l'espace H, c.-à-d. qu'elle satisfait aux conditions suivantes:

$$p(h) \geqslant 0, \quad \sum_{h \in H} p(h) = 1. \tag{8-1}$$

Désignons par g une certaine variante du jeu, c.-à-d. une des parties possibles. Cette variante sera définie si les stratégies x et y

des joueurs de même que la stratégie h des coups aléatoires sont choisies. Il s'ensuit que la partie concrète g représente un triplet des grandeurs x, y et h:

$$g = (x, y, h). \tag{8-2}$$

Le résultat de la partie sera exprimé par le gain ou la perte de chacun des joueurs. Pour plus de commodité, les gains et les pertes seront exprimés par un certain nombre, par exemple, par une somme d'argent en roubles.

Examinons une des parties concrètes $g(x, y, h)$ et notons respectivement par $L_x(x, y, h)$ et $L_y(x, y, h)$ les pertes du premier et du deuxième joueur, les gains étant considérés comme des pertes négatives. La somme totale des pertes des deux joueurs sera:

$$L_x(x, y, h) + L_y(x, y, h). \tag{8-3}$$

Dans ce qui suit, nous ne considérerons que les *jeux à somme nulle*, c.-à-d. les jeux dont la somme totale des pertes (8-3) est nulle. Dans ce cas, la perte de l'un des joueurs est égale au gain de l'autre.

En étudiant les jeux à somme nulle, il est superflu de comptabiliser séparément les pertes ou les gains des deux joueurs, car on peut se limiter seulement à l'examen de la perte du deuxième joueur (du gain du premier joueur):

$$L_y(x, y, h) = -L_x(x, y, h) = L(x, y, h). \tag{8-4}$$

Etant donné que la stratégie h est aléatoire, on voit que lorsqu'on choisit les stratégies x et y, la perte $L(x, y, h)$ sera une variable aléatoire avec la distribution de probabilités $p(h)$ sur l'espace H. C'est pourquoi on ne peut estimer les stratégies x et y choisies qu'en prenant la moyenne des pertes $L(x, y, h)$ sur tout l'espace H, c.-à-d. en introduisant la notion de pertes moyennes $L(x, y)$ déterminées conformément à (5-64) à partir de la relation

$$L(x, y) = \sum_{h \in H} L(x, y, h)\, p(h). \tag{8-5}$$

Le jeu sera défini si toutes les stratégies possibles des joueurs sont énumérées, c.-à-d. si les espaces X et Y sont donnés, et si pour tous $x \in X$ et $y \in Y$ les pertes $L(x, y)$ sont déterminées. De cette façon, nous arrivons à la définition formelle suivante du jeu.

Le jeu G est défini par le triplet

$$G = (X, Y, L), \tag{8-6}$$

où X et Y représentent certains espaces et L est une fonction numérique bornée définie sur le produit direct $X \times Y$. Les points $x \in X$ et $y \in Y$ sont appelés *stratégies* du premier et du deuxième joueur, tandis que la fonction L est nommée *fonction de pertes* (ou fonction de paiement).

Il est commode de représenter les jeux où chaque joueur dispose d'un nombre fini de stratégies (jeux finis) sous la forme d'une matrice dite de paiement. Soit $G = (X, Y, L)$ un jeu fini, où $X = \{x_1, \ldots, x_m\}$ et $Y = \{y_1, \ldots, y_n\}$. Alors, la matrice d'ordre $m \times n$

$$Q = \| q_{ij} \| = \begin{Vmatrix} q_{11} & \ldots & q_{1n} \\ \cdot & \cdot \cdot \cdot \cdot & \cdot \\ q_{m1} & \ldots & q_{mn} \end{Vmatrix} \tag{8-7}$$

dans laquelle $q_{ij} = L(x_i, y_j)$ est appelée *matrice du jeu G*.

Pour achever la description du jeu, il faut indiquer les buts qui conduisent les joueurs à choisir leurs stratégies. Ces buts sont assez simples. Le premier joueur tâche d'obtenir le gain maximal, c.-à-d. de maximiser la fonction $L(x, y)$, tandis que le deuxième joueur s'efforce de rendre sa perte minimale, c.-à-d. de minimiser la fonction $L(x, y)$. On voit donc que les buts des joueurs sont diamétralement opposés. Dans ce cas, la difficulté spécifique consiste dans le fait qu'aucun des joueurs n'a le contrôle absolu de la valeur de $L(x, y)$, car le premier ne dispose que des valeurs de x, tandis que le deuxième ne contrôle que les valeurs de y. L'élimination de cette difficulté, c.-à-d. la définition de la méthode la plus rationnelle de conduite du jeu par chacun des joueurs, constitue l'essentiel de la théorie des jeux.

Il faut souligner que les raisonnements ci-dessus ne sont valables que pour un jeu à deux personnes à somme nulle. Si au jeu participent plus de deux personnes, on est en présence d'une situation tout à fait nouvelle caractérisée par le fait que certains joueurs peuvent conjuguer leurs efforts dans le cadre d'une coalition qui prévoit la distribution contractuelle du gain réalisé. Les membres de la coalition peuvent estimer leurs possibilités d'une façon plus complète et entreprendre des actions concertées pour s'assurer des gains plus importants.

Une autre variante du jeu est le jeu à somme non nulle. Lorsqu'il s'agit d'un jeu pareil, les gains de certains joueurs peuvent être obtenus non seulement pour le compte des pertes des autres joueurs, mais aussi grâce à certains paiements versés de l'extérieur. Ces paiements peuvent être considérés comme la perte d'un joueur fictif supplémentaire, ce qui permet de ramener un jeu à n personnes à somme non nulle à un jeu à $n + 1$ personnes à somme nulle.

Vu que la théorie des jeux à n personnes, où $n > 2$, est assez compliquée et que son élaboration est loin d'être achevée, nous allons nous limiter à l'étude du jeu à deux personnes à somme nulle.

Exemple 8-1. Pour expliquer les notions introduites, examinons un jeu comportant quatre coups.

Premier coup (personnel). Le premier joueur choisit un des deux nombres entiers 1, 2.

Deuxième coup (aléatoire). On jette une pièce de monnaie et si, et seulement si, ce jet amène pile, on communique au deuxième joueur le choix fait par le premier.

Troisième coup (personnel). Le deuxième joueur choisit un des deux nombres entiers 3, 4.

Quatrième coup (aléatoire). On choisit au hasard avec les probabilités 0,4; 0,2; 0,4 un des trois nombres entiers 1, 2, 3.

Résultat du jeu: les nombres choisis aux premier, troisième et quatrième coups sont additionnés et la somme ainsi obtenue est payée par le deuxième joueur au premier si elle est paire, ou bien par le premier joueur au deuxième si elle est impaire.

Pour entreprendre l'analyse préliminaire du jeu, il est commode de le représenter sous la forme d'un arbre descendant dont les sommets représentent les situations qui surgissent pendant le déroulement

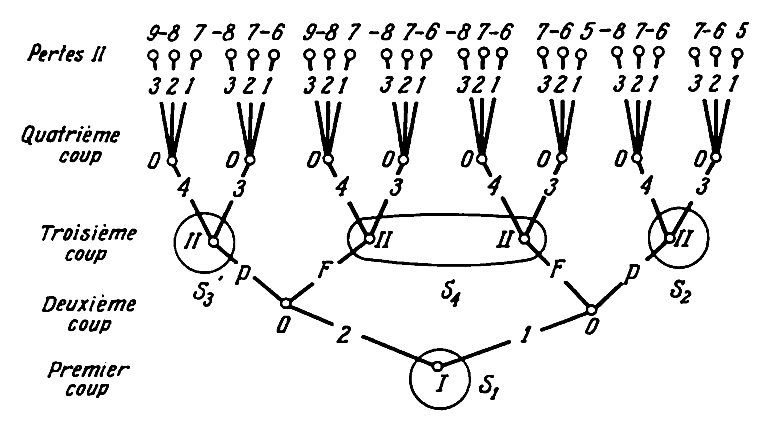


Fig. 8-1. Arbre du jeu

du jeu, tandis que les branches réunissant ces sommets correspondent aux coups joués. L'arbre du jeu qui nous intéresse est donné sur la figure 8-1. Les sommets correspondant aux coups personnels du premier et du deuxième joueur sont respectivement désignés par *I* et *II*. Les sommets correspondant aux coups aléatoires sont notés par *O*. Les sommets terminaux qui définissent les différentes variantes du jeu sont marqués de chiffres indiquant les pertes du deuxième joueur.

Dans différents sommets correspondant aux coups personnels le joueur possède une certaine information sur les coups précédents. Si dans certains sommets le joueur est en possession de la même information, il est commode de réunir ces sommets en obtenant ainsi des groupes de sommets S_i dits *classes d'information*. L'exemple examiné présente quatre classes d'information dont le contenu est le suivant:

S_1: il n'y a pas eu de coups; le premier joueur doit jouer le premier coup;

S_2: le premier joueur a choisi 1;

S_3: le premier joueur a choisi 2;

S_4: on ne sait pas le choix fait par le premier joueur.

Si le deuxième joueur tombe sur un sommet de la classe S_4, il est privé d'information sur le choix fait par le premier joueur. Un tel jeu est appelé *jeu avec information incomplète*. Si la classe d'information ne comporte qu'un seul sommet, le joueur qui y tombe est parfaitement informé sur tous les coups précédents (il peut obtenir cette information en suivant, par exemple, sur l'arbre du jeu le cheminement qui mène au sommet donné), c.-à-d. il est en possession de toute l'information concernant le jeu. Les jeux dont chaque classe d'information ne comporte qu'un seul sommet sont appelés *jeux avec information complète* (échecs, dames, dominos, etc.).

Examinons les espaces des stratégies des joueurs. L'espace des stratégies du premier joueur, qui ne comprend que deux éléments auxquels correspond le choix 1 ou 2, est décrit par le tableau 8-1, *a*. La stratégie du deuxième joueur doit indiquer son coup pour n'importe quelle variante de jeu, tandis que cette variante est définie par la classe d'information du joueur. Pour le deuxième joueur, on a trois classes d'information: S_2, S_3 et S_4. Ainsi donc la stratégie du deuxième joueur est d'indiquer lequel des deux nombres 3 ou 4

Tableau 8-1

Espaces des stratégies pour le jeu de l'exemple 8-1

x_1	x_2
(1)	(2)

a)

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
(333)	(334)	(343)	(344)	(433)	(434)	(443)	(444)

b)

h	$(P, 1)$	$(P, 2)$	$(P, 3)$	$(F, 1)$	$(F, 2)$	$(F, 3)$
$p(h)$	0,2	0,1	0,2	0,2	0,1	0,2

c)

il va choisir dans chaque classe d'information. La stratégie (4, 3, 3) signifie que le deuxième joueur choisit 4 dans la classe d'information S_2 et 3 dans les classes d'information S_3 et S_4. L'espace des stratégies du deuxième joueur est donné au tableau 8-1,*b*.

Les stratégies aléatoires h comportent deux coups: le deuxième coup (le choix de P ou de F avec les probabilités 0,5; 0,5) et le quatrième coup (le choix du nombre 1, 2 ou 3 avec les probabilités 0,4; 0,2; 0,4). La probabilité de chaque stratégie est égale au produit des probabilités des issues de ces deux coups. L'espace des stratégies aléatoires H et la distribution de probabilités $p(h)$ sont donnés au tableau 8-1,*c*.

Pour composer la matrice de paiement, il faut déterminer les pertes $q_{ij} = L(x_i, y_j)$ conformément à (8-5):

$$q_{ij} = \sum_{h \in H} L(x_i, y_j, h)\, p(h).$$

où les grandeurs $L(x_i, y_j, h)$ représentent les chiffres dont sont marqués les sommets terminaux de l'arbre du jeu. Les grandeurs q_{ij} sont commodes à calculer à l'aide des tableaux analogues au tableau 8-2 compilé pour la détermination de q_{11}.

Tableau 8-2

Calcul des pertes q_{11} pour le jeu de l'exemple 8-1

h	$p(h)$	$L(x_1, y_1, h)$	$L(x_1, y_1, h)\,p(h)$
P, 1	0,2	+5	+1,0
P, 2	0,1	−6	−0,6
P, 3	0,2	+7	+1,4
F, 1	0,2	+5	+1,0
F, 2	0,1	−6	−0,6
F, 3	0,2	+7	+1,4
	$q_{11} = +3,6$		

d) Valeurs supérieure et inférieure du jeu

Pour comprendre les principes qui sont à la base du choix que fait chaque joueur pour adopter sa stratégie, examinons le jeu dont la matrice est donnée par le tableau 8-3.

Supposons que le premier joueur ait choisi la stratégie x_k. Si le gain $L(x_k, y)$ dépend de la stratégie qui sera choisie par le deuxième joueur (par exemple, pour la stratégie x_1, les gains du premier joueur peuvent être 7, 2, 5, 1), le premier joueur pourra-t-il compter sur le gain maximal 7? Oui, s'il suppose que le deuxième joueur

Tableau 8-3

Matrice d'un jeu avec point-selle

	y_1	y_2	y_3	y_4	A (x)
x_1	7	2	5	1	1
x_2	2	2	3	4	2
x_3	5	3	4	4	3*
x_4	3	2	1	6	1
B (y)	7	3*	5	6	

choisira la stratégie y_1. Mais le deuxième joueur peut choisir n'importe quelle autre stratégie y compris y_4 pour laquelle le gain du premier joueur sera 1. Mais quelle que soit la stratégie du deuxième joueur, le gain du premier joueur ne sera jamais inférieur à 1. C'est pourquoi 1, qui est l'élément minimal de l'ensemble $L(x_1, y) = \{7, 2, 5, 1\}$, représente le gain garanti du premier joueur pour la stratégie x_1.

En généralisant les raisonnements ci-dessus, on voit que si le premier joueur adopte la stratégie x_k, il s'assurera un gain garanti A (x_k) égal à l'élément minimal de l'ensemble $L(x_k, y)$:

$$\mathrm{A}(x_k) = \min_{y} L(x_k, y). \tag{8-8}$$

La théorie des jeux suppose que les joueurs sont assez prudents et n'encourent pas de risques injustifiés. Dans ce cas, le premier joueur doit choisir la stratégie $x \in X$ qui correspond au nombre A (x) maximal. En désignant le gain garanti du premier joueur par α et en l'appelant *valeur nette inférieure du jeu*, on a

$$\alpha = \max_{x} \mathrm{A}(x) = \max_{x} \min_{y} L(x, y). \tag{8-9}$$

Les valeurs de A (x) correspondant à un jeu dont la matrice est de la forme donnée au tableau 8-3 sont portées dans la dernière colonne, dans laquelle la valeur de α est marquée d'un astérisque.

Des raisonnements analogues sont valables pour le deuxième joueur, à la différence que dans la matrice du jeu on indique ses pertes qu'il tend à rendre minimales. Examinons la stratégie y_h. Cette stratégie peut lui apporter une perte non supérieure à

$$\mathrm{B}(y_k) = \max_{x} L(x, y_k). \tag{8-10}$$

Pour s'assurer la perte minimale, le deuxième joueur doit adopter la stratégie $y \in Y$ qui correspond au nombre B (y) minimal. En désignant par β la perte à laquelle peut se limiter le deuxième joueur et en l'appelant *valeur nette supérieure du jeu*, on a

$$\beta = \min_y B(y) = \min_y \max_x L(x, y). \qquad (8\text{-}11)$$

Les valeurs de B (y) pour un jeu dont la matrice est de la forme donnée au tableau 8-3 sont portées dans la ligne inférieure du tableau, dans laquelle la valeur de β est marquée d'un astérisque.

Théorème 8-1. *Si* $G = (X, Y, L)$ *est un certain jeu, pour tous* $x \in X$ *et* $y \in Y$, *on a*

$$A(x) \leqslant L(x, y) \leqslant B(y) \text{ et } \alpha \leqslant \beta.$$

Démonstration. Par définition

$$A(x) = \min_y L(x, y) \leqslant L(x, y); \qquad (8\text{-}12)$$

$$B(y) = \max_x L(x, y) \geqslant L(x, y). \qquad (8\text{-}13)$$

Donc,

$$A(x) \leqslant L(x, y) \leqslant B(y). \qquad (8\text{-}14)$$

Etant donné que cette relation est vérifiée pour n'importe quels $x \in X$ et $y \in Y$, en choisissant pour x la valeur pour laquelle A (x) = $= \alpha$ et pour y la valeur pour laquelle B $(y) = \beta$, on obtient:

$$\alpha \leqslant \beta. \qquad (8\text{-}15)$$

On voit donc que la valeur inférieure du jeu, c.-à-d. le gain que peut s'assurer le premier joueur, ne dépasse pas la valeur supérieure du jeu, c.-à-d. la perte à laquelle peut se borner le deuxième joueur.

8-2. VALEURS ET STRATÉGIES OPTIMALES DES JEUX

a) Jeu avec point-selle

Le cas particulier le plus simple de jeu se présente lorsque $\alpha = \beta$. Désignons cette grandeur par c. C'est justement à ce cas que se rapporte le jeu dont la matrice est donnée au tableau 8-3, où $\alpha = \beta = 3$.

Ici, aucune stratégie ne peut garantir au premier joueur un gain supérieur à β, car c'est justement à cette grandeur β que le deuxième joueur peut limiter sa perte. D'autre part, aucune stratégie ne peut garantir au deuxième joueur une perte inférieure à α, étant donné que le premier joueur peut s'assurer un gain égal à α. Il s'ensuit donc que si $\alpha = \beta = c$, aucune stratégie des deux joueurs ne peut leur garantir un résultat meilleur que c. En même temps, chacun des joueurs peut s'assurer le résultat c. Autrement dit, ni le premier, ni le deuxième joueur ne dispose d'une stratégie meilleure que celle

qui leur assure le résultat c. Dans ce cas, c est appelé *valeur nette du jeu*, tandis que les stratégies des joueurs qui leur assurent le résultat c sont dites *stratégies optimales*.

La case de la matrice qui définit la grandeur c est appelée *point-selle*, car la valeur de c est le maximum de la colonne et le minimum de la ligne à l'intersection desquelles se trouve cette grandeur. Pour cette raison, un jeu dont la valeur est nette est appelé *jeu avec point-selle*.

Un jeu avec point-selle est dit *équitable* si $c = 0$. Si $c \neq 0$, le jeu est non équitable et pour le rendre équitable, en début de chaque nouvelle partie le premier joueur doit payer la quantité c au deuxième joueur.

b) Stratégies pures et mixtes

Si le jeu n'a pas de point-selle, la détermination de la valeur du jeu et des stratégies optimales des joueurs devient plus difficile. Examinons, par exemple, le jeu dont la matrice est donnée au tableau 8-4. Dans ce jeu, $\alpha = 4$ et $\beta = 5$. On voit donc que le premier joueur peut s'assurer un gain égal à 4 tandis que le deuxième joueur peut limiter sa perte à 5. Le domaine situé entre β et α n'appartient à personne et chacun des joueurs peut tâcher d'améliorer son résultat pour le compte de ce domaine. Quelles doivent être, dans ce cas, les stratégies optimales des joueurs?

Tableau 8-4

Matrice d'un jeu sans point-selle

	y_1	y_2	A (x)
x_1	3	6	3
x_2	5	4	4*
B (y)	5*	6	

Si chacun des joueurs adopte la stratégie marquée d'un astérisque (x_2 et y_1), le gain du premier joueur et la perte du deuxième seront égaux à 5. Cela ne convient pas au deuxième joueur, car le premier gagne plus qu'il ne peut s'assurer. Mais si le deuxième joueur arrive d'une façon ou d'une autre à deviner les intentions du premier joueur relatives à l'adoption de la stratégie x_2, il pourra alors adopter la stratégie y_2 en réduisant ainsi à 4 le gain du premier joueur. Il est vrai que si le premier joueur apprend les dessins du deuxième quant à l'adoption de la stratégie y_2, il pourra mettre en œuvre la stratégie x_1 en portant ainsi son gain à 6. De cette façon, il se crée une situation où chaque joueur doit tenir en secret la stratégie qu'il a l'intention d'adopter. Mais comment y parvenir? Si la partie est réitérée plusieurs fois et si le deuxième joueur adopte toujours la stratégie y_2, le premier joueur arrivera bientôt à deviner ses intentions et, en adoptant la

stratégie x_1, il obtiendra un gain supplémentaire. Il est évident que le deuxième joueur doit changer de stratégie dans chaque nouvelle partie, mais ses actions doivent être telles que le premier joueur ne devine pas la stratégie qu'il va adopter chaque fois.

Le secret peut être gardé si chaque fois la stratégie est choisie au hasard en utilisant à cette fin un certain mécanisme aléatoire. Par exemple, le deuxième joueur peut jeter une pièce de monnaie et adopter la stratégie y_1 si le jet amène pile, ou la stratégie y_2 si c'est face. Cette façon d'agir prive l'adversaire de toute possibilité d'anticiper les actions de l'autre joueur.

Lorsqu'on utilise un mécanisme aléatoire, les gains et les pertes des joueurs seront des variables aléatoires. Dans ce cas, le résultat du jeu peut être apprécié par la perte moyenne du deuxième joueur. Ainsi, lorsque dans un jeu dont la matrice est de la forme donnée par le tableau 8-4 le deuxième joueur adopte les stratégies y_1 et y_2 d'une façon aléatoire avec les probabilités 0,5 ; 0,5, sa perte moyenne, pour la stratégie x_1 du premier joueur, sera :

$$L_m = q_{11} \cdot 0{,}5 + q_{12} \cdot 0{,}5 = 3 \cdot 0{,}5 + 6 \cdot 0{,}5 = 4{,}5,$$

et pour la stratégie x_2 du premier joueur

$$L_m = q_{21} \cdot 0{,}5 + q_{22} \cdot 0{,}5 = 5 \cdot 0{,}5 + 4 \cdot 0{,}5 = 4{,}5.$$

Donc, le deuxième joueur peut limiter sa perte moyenne à 4,5 indépendamment de la stratégie adoptée par le premier joueur.

On voit de cette façon que dans certains cas il n'est pas rationnel de désigner à l'avance la stratégie qui sera adoptée et qu'il est préférable de la choisir au hasard en utilisant à cette fin un mécanisme aléatoire. La stratégie basée sur un choix aléatoire sera appelée *stratégie mixte* contrairement aux stratégies désignées à l'avance et examinées précédemment qui seront maintenant appelées *stratégies pures.*

Donnons maintenant une définition plus rigoureuse des stratégies pures et mixtes.

Soit $G = (X, Y, L)$ un certain jeu. Les espaces $X = \{x_1, \ldots, x_m\}$ et $Y = \{y_1, \ldots, y_n\}$, qui comportent les énumérations de toutes les stratégies possibles des joueurs, sont appelés *espaces des stratégies pures*, les éléments $x \in X$ et $y \in Y$ de ces espaces étant les stratégies pures des joueurs.

Pour obtenir une stratégie mixte, le joueur doit utiliser un certain mécanisme aléatoire (jet d'une pièce de monnaie, d'un dé, etc.) dont le nombre d'issues est égal au nombre de stratégies pures se trouvant à sa disposition.

Supposons que le mécanisme aléatoire du premier joueur est à m issues qui forment l'ensemble $R = \{r^{(1)}, \ldots, r^{(m)}\}$. Désignons par $\xi^{(1)}, \ldots, \xi^{(m)}$ les probabilités avec lesquelles apparaissent les issues correspondantes du mécanisme aléatoire.

16—0732

La stratégie mixte du premier joueur consiste dans le fait qu'à chaque issue $r \in R$ on associe une stratégie pure $x \in X$. De cette façon les grandeurs $\xi^{(1)}, \ldots, \xi^{(m)}$ représenteront les probabilités avec lesquelles sont utilisées les stratégies pures $x_1, \ldots, x_m$. L'ensemble ordonné $\xi = (\xi^{(1)}, \ldots, \xi^{(m)})$ dont les éléments satisfont aux conditions

$$\xi^{(i)} \geqslant 0, \quad \sum_i \xi^{(i)} = 1 \tag{8-16}$$

peut maintenant être considéré en tant que distribution de probabilités $\xi(x)$ sur l'espace X. Cette distribution de probabilités définit complètement le caractère du jeu du premier joueur et est appelée sa *stratégie mixte* correspondant au mécanisme aléatoire donné.

Un autre mécanisme aléatoire amènera à une autre distribution de probabilités $\xi'(x)$ et déterminera une autre stratégie mixte du premier joueur.

Dans le cas général, le premier joueur peut disposer d'une infinité de mécanismes aléatoires déterminant toutes les distributions de probabilités possibles sur l'espace de ses stratégies pures:

$$\xi_1 = (\xi_1^{(1)}, \ldots, \xi_1^{(m)}); \quad \xi_2 = (\xi_2^{(1)}, \ldots, \xi_2^{(m)}) \ldots \tag{8-17}$$

Dans ce cas, l'ensemble

$$\mathrm{E} = \{\xi_1, \xi_2, \ldots\} \tag{8-18}$$

représentera l'*espace des stratégies mixtes du premier joueur.*

D'une façon analogue, le deuxième joueur peut utiliser son mécanisme aléatoire qui définit les probabilités $\eta^{(1)}, \ldots, \eta^{(n)}$ avec lesquelles seront utilisées les stratégies pures $y_1, \ldots, y_n$. Dans ce cas, l'ensemble ordonné $\eta = (\eta^{(1)}, \ldots, \eta^{(n)})$ dont les éléments vérifient les relations

$$\eta^{(k)} \geqslant 0, \quad \sum_k \eta^{(k)} = 1 \tag{8-19}$$

représente la distribution de probabilités $\eta(y)$ sur l'espace Y et est appelé *stratégie mixte* du deuxième joueur.

Le deuxième joueur, aussi bien que le premier, peut disposer d'une infinité de mécanismes aléatoires déterminant les différentes distributions de probabilités sur l'espace de ses stratégies pures:

$$\eta_1 = (\eta_1^{(1)}, \ldots, \eta_1^{(n)}); \quad \eta_2 = (\eta_2^{(1)}, \ldots, \eta_2^{(n)}) \ldots, \tag{8-20}$$

dont l'ensemble

$$\mathrm{H} = \{\eta_1, \eta_2, \ldots\} \tag{8-21}$$

forme l'*espace des stratégies mixtes du deuxième joueur.*

c) Fonction de pertes lors de l'utilisation des stratégies mixtes

Examinons le jeu $G = (X, Y, L)$ dans lequel $X = \{x_1, \ldots, x_m\}$ et $Y = \{y_1, \ldots, y_n\}$ sont les espaces des stratégies pures des joueurs. $L(x, y)$ représentant les pertes du deuxième joueur déterminées pour les stratégies pures $x \in X$ et $y \in Y$. Dans ce qui suit, la fonction $L(x, y)$ sera appelée *fonction de pertes*.

Supposons maintenant que les joueurs adoptent des stratégies mixtes. Cela signifie que l'on donne les ensembles $\mathrm{E} = \{\xi_1, \xi_2, \ldots\}$ et $\mathrm{H} = \{\eta_1, \eta_2, \ldots\}$ dont les éléments représentent les stratégies mixtes des joueurs, c.-à-d. les différentes distributions de probabilités $\xi(x)$ et $\eta(y)$ sur les espaces X et Y. Le caractère du jeu sera défini lorsque chaque joueur choisira sa stratégie mixte $\xi \in \mathrm{E}$ et $\eta \in \mathrm{H}$. Il s'ensuit que, lors de l'utilisation des stratégies mixtes, le jeu n'est plus défini par les ensembles X et Y mais par les ensembles E et H.

Lorsqu'on adopte des stratégies mixtes, la perte du deuxième joueur, donc la fonction de pertes, subira des changements. Etant donné que le caractère du jeu devient aléatoire, les gains et les pertes des joueurs revêtent aussi un caractère aléatoire. Pour cette raison, on ne peut parler maintenant que de la valeur moyenne du gain X ou de la perte Y, valeur déterminée aussi bien par la fonction de pertes $L(x, y)$ que par les distributions de probabilités $\xi(x)$ et $\eta(y)$ et qui peut être trouvée d'après la formule de la valeur moyenne d'une fonction de deux variables:

$$L(\xi, \eta) = \sum_{x, y} L(x, y)\, \xi(x)\, \eta(y). \tag{8-22}$$

La fonction de pertes $L(\xi, \eta)$ doit être présente dans la définition du jeu lors de l'utilisation des stratégies mixtes.

De cette façon, lorsqu'on utilise des stratégies mixtes, à la place du jeu $G = (X, Y, L)$ on obtient un nouveau jeu $\Gamma = (\mathrm{E}, \mathrm{H}, L)$ qui est le centrage du jeu G. Dans ce nouveau jeu, la fonction de pertes $L(\xi, \eta)$ est déterminée suivant la formule (8-22) en tant que moyenne de $L(x, y)$ pour les distributions de probabilités $\xi(x)$ et $\eta(y)$ données.

Tout ce qui vient d'être dit peut être résumé comme suit.

Soit $G = (X, Y, L)$ un certain jeu et supposons que E et H définissent les différentes distributions de probabilités ξ et η sur les espaces X et Y, $L(\xi, \eta)$ étant la valeur moyenne de la fonction de pertes. Alors le jeu

$$\Gamma = (\mathrm{E}, \mathrm{H}, L) \tag{8-23}$$

est le *centrage* du jeu G.

Soit $\Gamma = (E, H, L)$ le centrage du jeu $G = (X, Y, L)$. Alors les points $x \in X$ et $y \in Y$ sont appelés *stratégies pures* dans le jeu G, les points $\xi \in E$ et $\eta \in H$ étant appelés *stratégies mixtes* dans le même jeu.

Remarquons que les stratégies pures peuvent être considérées comme un cas particulier des stratégies mixtes. En effet, la stratégie mixte définie par la distribution de probabilités

$$\xi(x) = \begin{cases} 1 & \text{pour} \quad x = x_k; \\ 0 & \text{pour} \quad x \neq x_k \end{cases} \tag{8-24}$$

coïncide avec la stratégie pure x_k, tandis que la stratégie mixte définie par la distribution de probabilités

$$\eta(y) = \begin{cases} 1 & \text{pour} \quad y = y_i; \\ 0 & \text{pour} \quad y \neq y_i \end{cases} \tag{8-25}$$

coïncide avec la stratégie pure y_i.

d) Valeurs supérieure et inférieure du jeu lors de l'utilisation des stratégies mixtes

Soient $G = (X, Y, L)$ un certain jeu et $\Gamma = (E, H, L)$ son centrage.

Supposons que le premier joueur adopte la stratégie mixte $\xi \in E$. Ce joueur veut savoir quel sera son gain garanti, c.-à-d. le gain minimal qu'il puisse s'assurer à coup sûr même si le deuxième joueur adopte sa meilleure stratégie mixte η. Le cas où le deuxième joueur adopte une stratégie pure quelconque n'est pas exclu.

Désignons par $A(\xi)$ le gain garanti du premier joueur pour la stratégie ξ. Il est évident que, pour la stratégie ξ donnée et pour différentes $\eta \in H$, c'est la limite inférieure de la fonction de gain $L(\xi, \eta)$, c.-à-d.

$$A_\Gamma(\xi) = \min_\eta L(\xi, \eta). \tag{8-26}$$

Ensuite, le premier joueur voudra savoir laquelle de toutes les stratégies possibles $\xi \in E$ lui rendra son gain garanti maximal, c.-à-d. la limite supérieure de la fonction $A_\Gamma(\xi)$ désignée par α_Γ et appelée *valeur inférieure du jeu* en présence de stratégies mixtes des joueurs:

$$\alpha_\Gamma = \max_\xi A_\Gamma(\xi) = \max_\xi \min_\eta L(\xi, \eta). \tag{8-27}$$

La stratégie pour laquelle la condition (8-27) est vérifiée est appelée stratégie *minimax* (parfois on dit maximin) du premier joueur et est désignée par ξ_0.

Supposons maintenant que le deuxième joueur ait choisi une certaine stratégie mixte $\eta \in$ H. Ce joueur veut savoir quelle sera sa perte maximale pour la meilleure stratégie $\xi \in$ E du premier joueur, c.-à-d. la limite supérieure de la fonction $L(\xi, \eta)$, pour η donné et différents $\xi \in$ E, limite désignée par $B_\Gamma(\eta)$:

$$B_\Gamma(\eta) = \max_\xi L(\xi, \eta). \tag{8-28}$$

Le deuxième joueur veut aussi savoir comment choisir une stratégie $\eta \in$ H pour laquelle sa perte soit minimale, c.-à-d. la limite inférieure de la fonction $B_\Gamma(\eta)$, désignée par β_Γ et appelée *valeur supérieure du jeu* en présence de stratégies mixtes des joueurs:

$$\beta_\Gamma = \min_\eta B_\Gamma(\eta) = \min_\eta \max_\xi L(\xi, \eta). \tag{8-29}$$

La stratégie pour laquelle la condition (8-29) est satisfaite est appelée stratégie *minimax* du deuxième joueur et se note η_0.

Pour faciliter les raisonnements ultérieurs, introduisons les désignations suivantes pour les valeurs inférieure et supérieure du jeu $G = (X, Y, L)$ lors de l'utilisation des stratégies pures:

$A_G(x) = \min_y L(x, y)$: gain garanti du premier joueur pour la stratégie $x \in X$;

$\alpha_G = \max_x A_G(x)$: valeur inférieure du jeu G en présence de stratégies pures des joueurs;

$B_G(y) = \max_x L(x, y)$: perte maximale du deuxième joueur pour la stratégie $y \in Y$;

$\beta_G = \min_y B_G(y)$: valeur supérieure du jeu G en présence de stratégies pures des joueurs.

Théorème 8-2. *Si l'on a un jeu $G = (X, Y, L)$ et si $\Gamma =$ = (E, H, L) est le centrage du jeu G, alors*

$$\left.\begin{array}{ll} \text{a)} & A_\Gamma(\xi) = \min_y L(\xi, y); \\ \text{b)} & \alpha_G \leqslant \alpha_\Gamma; \\ \text{c)} & B_\Gamma(\eta) = \max_x L(x, \eta); \\ \text{d)} & \beta_G \geqslant \beta_\Gamma; \\ \text{e)} & \alpha_\Gamma \leqslant \beta_\Gamma. \end{array}\right\} \tag{8-30}$$

Le point a) nous dit que si le premier joueur choisit n'importe quelle stratégie mixte, son gain garanti sera égal au gain garanti lors de l'utilisation par le deuxième joueur seulement des stratégies pures. Conformément au point b), la valeur inférieure du jeu en présence de stratégies mixtes du premier joueur n'est pas inférieure à la valeur inférieure du jeu pour les stratégies pures, c.-à-d. il y a

une stratégie mixte qui, en tout cas, n'est pas pire que la stratégie pure optimale. Les points c) et d) contiennent des assertions analogues relatives au deuxième joueur. Le point e) signifie que la valeur inférieure du jeu α_Γ, lors de l'utilisation des stratégies mixtes, ne dépasse pas la valeur supérieure du jeu, c.-à-d. qu'en adoptant les meilleures stratégies mixtes, le gain garanti du premier joueur ne sera pas supérieur à la perte assurée du deuxième joueur.

D é m o n s t r a t i o n. a) Etant donné que les stratégies pures $y \in Y$ sont des cas particuliers des stratégies mixtes $\eta \in H$, $L(\xi, y)$ est un cas particulier de $L(\xi, \eta)$, donc un sous-ensemble de $L(\xi, \eta)$. En vertu du théorème de la borne inférieure d'un sous-ensemble (voir paragraphe 1-1), on a:

$$\min_y L(\xi, y) \geqslant \min_\eta L(\xi, \eta) = A_\Gamma(\xi). \tag{8-31}$$

Mais $L(\xi, \eta)$ peut être considérée comme la moyenne de $L(\xi, y)$ prise sur tous les y avec la distribution de probabilités $\eta(y)$. En vertu du théorème de la moyenne (voir paragraphe 5-5), on obtient:

$$\min_y L(\xi, y) \leqslant L(\xi, \eta). \tag{8-32}$$

La relation (8-32) reste vraie pour toute $\eta \in H$, y compris η_0, pour laquelle $L(\xi, \eta_0) = \min_\eta L(\xi, \eta)$. De cette façon,

$$\min_y L(\xi, y) \leqslant \min_\eta L(\xi, \eta) = A_\Gamma(\xi). \tag{8-33}$$

En comparant (8-31) à (8-33), on peut s'assurer que le point a) est vrai.

b) En comparant $A_G(x) = \min_y L(x, y)$ avec la grandeur $A_\Gamma(\xi)$, qui en vertu du point a) peut être définie de la même manière que $A_G(x)$ seulement pour les stratégies pures y du deuxième joueur, et tenant compte du fait que la stratégie pure x est un cas particulier de la stratégie mixte $\xi(x)$, nous voyons que $A_G(x)$ est le sous-ensemble de $A_\Gamma(\xi)$. En vertu du théorème de la borne supérieure d'un sous-ensemble, on a:

$$\max_x A_G(x) \leqslant \max_\xi A_\Gamma(\xi), \tag{8-34}$$

ce qui démontre le point b).

Les points c) et d) sont démontrés d'une façon analogue.

e) Par définition, on a:

$$\left.\begin{aligned} A_\Gamma(\xi) &= \min_\eta L(\xi, \eta) \leqslant L(\xi, \eta), \\ B_\Gamma(\eta) &= \max_\xi L(\xi, \eta) \geqslant L(\xi, \eta). \end{aligned}\right\} \tag{8-35}$$

Donc

$$A_\Gamma(\xi) \leqslant L(\xi, \eta) \leqslant B_\Gamma(\eta). \tag{8-36}$$

Cela reste vrai pour toutes ξ et η, c.-à-d. pour une ξ telle que $A_\Gamma(\xi) = \alpha_\Gamma$ et pour une η telle que $B_\Gamma(\eta) = \beta_\Gamma$, ce qui démontre le point e).

Corollaire. En confrontant entre eux les points b), d) et e), on trouve:

$$\alpha_G \leqslant \alpha_\Gamma \leqslant \beta_\Gamma \leqslant \beta_G. \tag{8-37}$$

De la relation (8-37) il découle que si le jeu G a une valeur nette, c.-à-d. si $\alpha_G = \beta_G = c$ (nous avons vu que cela a lieu pour un jeu avec point-selle), alors $\alpha_\Gamma = \beta_\Gamma = c$, c.-à-d. le jeu Γ a lui aussi une valeur nette. C'est pourquoi la stratégie optimale dans le jeu G est aussi la stratégie optimale dans le jeu Γ. Dans ce cas, il est superflu d'examiner le jeu Γ, et les stratégies optimales des joueurs peuvent être trouvées par la méthode décrite pour le jeu avec point-selle.

Mais le jeu Γ a été introduit dans le but d'analyser les jeux pour lesquels $\alpha_G < \beta_G$, et on a le droit de poser la question suivante: n'est-il pas vrai que, dans ce cas, pour une grande classe de jeux, $\alpha_\Gamma = \beta_\Gamma = c_\Gamma$, et n'est-il pas permis de considérer c_Γ en tant que valeur nette du jeu Γ? Dans ce qui suit nous allons montrer qu'effectivement cela a lieu, de façon que l'adoption des stratégies mixtes permet, dans ce cas aussi, de trouver la valeur du jeu et les stratégies optimales des joueurs. Donnons pour cela une autre définition de la stratégie optimale.

Soient un jeu $G = (X, Y, L)$ et son centrage $\Gamma = (\mathrm{E}, \mathrm{H}, L)$. Si le jeu Γ a une valeur nette $c_\Gamma = \alpha_\Gamma = \beta_\Gamma$, on dit que la valeur du jeu G est c_Γ, et toute stratégie optimale dans le jeu Γ, c.-à-d. une stratégie qui assure un gain garanti du premier joueur et une perte garantie du deuxième joueur égaux à c_Γ, est appelée stratégie *optimale* dans le jeu G. Il est aisé de voir que si $\alpha_\Gamma = \beta_\Gamma = c_\Gamma$, les stratégies minimax $\xi_0(x)$ et $\eta_0(y)$ des joueurs seront leurs stratégies optimales.

Le théorème qui affirme que chaque jeu fini a une valeur et que chaque joueur a à sa disposition des stratégies optimales est le théorème fondamental de la théorie des jeux. La démonstration de ce théorème fera l'objet du paragraphe suivant.

8-3. THÉORÈME FONDAMENTAL DE LA THÉORIE DES JEUX

a) S-jeu

Une interprétation géométrique utile peut être donnée aux jeux dans lesquels le premier joueur dispose d'un nombre fini de stratégies pures. Soit un jeu $G = (X, Y, L)$ dont la matrice $Q = \| q_{ij} \|$ est définie par l'expression (8-7). Associons à chaque $y \in Y$ un point C dans un espace à m dimensions de façon que ses coordonnées représentent les pertes du deuxième joueur pour toutes les stratégies

pures possibles du premier joueur:

$$C = [L(x_1, y), \ldots, L(x_m, y)]. \tag{8-38}$$

Tenant compte du fait que $L(x_i, y_j) = q_{ij}$ et en donnant à y toutes les valeurs possibles, nous obtiendrons n points:

$$C_1 = (q_{11}, \ldots, q_{m1}); \ldots; C_n = (q_{1n}, \ldots, q_{mn}), \tag{8-39}$$

dont les coordonnées sont représentées par les pertes données par les colonnes correspondantes de la matrice du jeu. Le jeu donné par l'ensemble des points

$$\{C_1, \ldots, C_n\}, \tag{8-40}$$

a été appelé *S-jeu* et se joue comme suit.

Le deuxième joueur choisit un des points C_i de l'ensemble (8-40), par exemple, le point C_3. Indépendamment du choix du deuxième joueur, le premier joueur choisit une coordonnée du point C_i, par exemple, la deuxième. Dans ce cas, la perte du deuxième joueur sera égale à la valeur de la deuxième coordonnée du point C_3, c.-à-d. q_{23}.

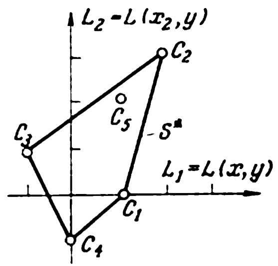


Fig. 8-2. Représentation géométrique d'un S-jeu

On voit donc qu'un S-jeu équivaut à un jeu ordinaire, tout au moins lorsqu'on adopte des stratégies pures, car le choix d'un point de l'ensemble (8.40) équivaut au choix de la stratégie $y \in Y$, tandis que le choix de la coordonnée de ce point équivaut au choix de la stratégie $x \in X$.

Si le premier joueur ne dispose que de deux stratégies, le S-jeu peut être représenté par une construction géométrique, car, dans ce cas, les points de l'ensemble (8-40) seront situés dans le plan.

Exemple 8-2. Examinons le jeu dont la matrice est donnée par le tableau 8-5. Le S-jeu équivalent comporte cinq points: $C_1 = (1, 0)$, $C_2 = (2, 3)$, $C_3 = (-1, 1)$, $C_4 = (0, -1)$, $C_5 = (1, 2)$. La figure 8-2 donne la représentation géométrique de ce jeu.

Tableau 8-5

	y_1	y_2	y_3	y_4	y_5
x_1	1	2	−1	0	1
x_2	0	3	1	−1	2

Désignons par S^* l'enveloppe convexe de l'ensemble fini des points $\{C_1, \ldots, C_n\}$ et démontrons le théorème qui dit qu'un S-jeu équivaut à un jeu ordinaire non seulement lorsqu'on utilise

des stratégies pures, mais aussi lorsque les stratégies adoptées sont mixtes.

Théorème 8-3. *Toute stratégie mixte du deuxième joueur peut être représentée par un point appartenant à l'enveloppe convexe S^*, et vice versa, tout point $S \in S^*$ peut être considéré comme une certaine stratégie mixte du deuxième joueur.*

Démonstration. Considérons les stratégies mixtes $\xi = (\xi^{(1)}, \ldots, \xi^{(m)})$ et $\eta = (\eta^{(1)}, \ldots, \eta^{(n)})$ des joueurs. Les pertes du deuxième joueur qui adopte les stratégies mixtes données seront:

$$L(\xi, \eta) = \sum_i \sum_k q_{ik}\xi^{(i)}\eta^{(k)} = \sum_i \xi^{(i)} \sum_k q_{ik}\eta^{(k)} = \sum_i \xi^{(i)} S^{(i)}, \qquad (8\text{-}41)$$

où

$$S^{(i)} = \sum_k q_{ik}\eta^{(k)}. \qquad (8\text{-}42)$$

Désignons par S un point de coordonnées $S^{(1)}, \ldots, S^{(m)}$ dans un espace à m dimensions où conformément à (8-42)

$$\left.\begin{array}{l} S^{(1)} = q_{11}\eta^{(1)} + \ldots + q_{1n}\eta^{(n)}; \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ S^{(m)} = q_{m1}\eta^{(1)} + \ldots + q_{mn}\eta^{(n)}. \end{array}\right\} \qquad (8\text{-}43)$$

Ayant en vue que $(q_{1i}, \ldots, q_{mi}) = C_i$, les expressions (8-43) peuvent être écrites sous la forme d'une seule relation vectorielle

$$S = C_1\eta^{(1)} + \ldots + C_n\eta^{(n)} = \sum_i C_i\eta^{(i)}. \qquad (8\text{-}44)$$

Compte tenu du fait que les grandeurs $\eta^{(i)}$ vérifient les relations (8-19), on voit que S n'est autre chose que la moyenne pondérée des points $C_1, \ldots, C_n$ respectivement de poids $\eta^{(1)}, \ldots, \eta^{(n)}$, donc S est un point qui appartient à l'enveloppe convexe S^*. De cette façon, à chaque stratégie $\eta = (\eta^{(1)}, \ldots, \eta^{(n)})$ du deuxième joueur correspondra un certain point appartenant à l'enveloppe convexe S^*, et la définition de ce point est équivalente à la définition de la stratégie mixte du deuxième joueur.

Le contraire est aussi vrai. Etant donné que tout point S appartenant à l'enveloppe convexe S^* peut être présenté comme la moyenne pondérée des points $C_1, \ldots, C_n$, définissant l'enveloppe convexe S^*, c.-à-d. peut être présenté sous la forme de l'expression (8-44), alors, pour chaque point $S \in S^*$, on trouvera des poids $\eta^{(1)}, \ldots, \eta^{(n)}$ tels que leur définition détermine la stratégie mixte du deuxième joueur.

Corollaire. Vu que la stratégie mixte du premier joueur reste dans un S-jeu la même que dans un jeu ordinaire, il découle du théorème qui vient d'être démontré qu'un S-jeu est tout à fait équivalent à un jeu ordinaire, c.-à-d. que n'importe quel jeu peut être présenté comme un S-jeu équivalent.

Dans ce qui suit, le S-jeu sera désigné par Γ_1. Pour passer du jeu $\Gamma = (E, H, L)$ au S-jeu, au lieu d'utiliser l'espace des stratégies mixtes du deuxième joueur $H = \{\eta_1, \eta_2, \ldots\}$ il faut faire appel à l'espace des S-stratégies, c.-à-d. à l'enveloppe convexe S^*. Si l'on désigne les pertes du deuxième joueur dans le S-jeu par L_1, le S-jeu sera défini par le triplet E, S^* et L_1, les pertes L_1 devant être définies sur le produit direct $E \times S^*$. De cette façon, l'expression

$$\Gamma_1 = (E, S^*, L_1) \tag{8-45}$$

définit le S-jeu. En vertu de (8-41), la fonction de pertes $L_1(\xi, S)$ sera définie comme suit:

$$L_1(\xi, S) = \sum_i \xi^{(i)} S^{(i)} = \xi S, \tag{8-46}$$

où par ξS on désigne le produit scalaire des vecteurs

$$\xi = (\xi^{(1)}, \ldots, \xi^{(m)}) \text{ et } S = (S^{(1)}, \ldots, S^{(m)}).$$

b) Valeurs inférieure et supérieure du jeu dans un S-jeu

Si le premier joueur adopte dans un S-jeu la stratégie mixte $\xi \in E$, son gain garanti sera:

$$A_{\Gamma_1}(\xi) = \min_S L_1(\xi, S) = \min_S \xi S. \tag{8-47}$$

Désignons par $\xi_0 = (\xi_0^{(1)}, \ldots, \xi_0^{(m)})$ la stratégie du premier joueur pour laquelle $A_{\Gamma_1}(\xi)$ atteint son maximum. La valeur de $A_{\Gamma_1}(\xi)$ sera égale à la valeur inférieure du jeu α_{Γ_1} qui, en vertu de l'équivalence d'un S-jeu à un jeu ordinaire, coïncide avec α_Γ. Donc

$$\alpha_\Gamma = A_{\Gamma_1}(\xi_0) = \max_\xi A_{\Gamma_1}(\xi) = \max_\xi \min_S \xi S. \tag{8-48}$$

La stratégie ξ_0 satisfaisant à la relation (8-48) est appelée stratégie *minimax* du premier joueur.

Supposons maintenant que le deuxième joueur adopte une certaine stratégie $S \in S^*$. Dans ce cas, la perte à laquelle il va se limiter sera

$$B_{\Gamma_1}(S) = \max_\xi L_1(\xi, S) = \max_\xi \xi S. \tag{8-49}$$

Désignons par $S_0 = (S_0^{(1)}, \ldots, S_0^{(m)})$ le point de l'enveloppe convexe S^* pour lequel la grandeur $B_{\Gamma_1}(S)$ atteint son minimum. La valeur de $B_{\Gamma_1}(S_0)$ sera égale à la valeur supérieure du jeu β_{Γ_1} qui, en vertu de l'équivalence d'un S-jeu à un jeu ordinaire, coïncide avec β_Γ. Donc

$$\beta_\Gamma = B_{\Gamma_1}(S_0) = \min_S B_{\Gamma_1}(S) = \min_S \max_\xi \xi S. \tag{8-50}$$

La stratégie S_0 satisfaisant à la relation (8-50) est appelée stratégie *minimax* du deuxième joueur.

Les expressions de $B_{\Gamma_1}(S)$ et de β_Γ peuvent être mises sous une forme plus commode à l'aide du théorème ci-après.

Théorème 8-4. *Si S est un point arbitraire d'un espace à m dimensions et si* $\xi = (\xi^{(1)}, \ldots, \xi^{(m)})$ *est une variable multidimensionnelle satisfaisant à la condition* (8-16), *ona la relation*

$$\max_\xi \xi S = \max(S^{(1)}, \ldots, S^{(m)}). \tag{8-51}$$

D é m o n s t r a t i o n. Soit $S^{(k)} = \max(S^{(1)}, \ldots, S^{(m)})$. Examinons la valeur particulière de ξ correspondant au cas suivant:

$$\xi^{(i)} = \begin{cases} 1 \text{ pour } i = k; \\ 0 \text{ pour } i \neq k. \end{cases} \tag{8-52}$$

Dans ce cas, $\xi S = S^{(k)}$. De cette façon, $S^{(k)}$ est une valeur particulière du produit scalaire ξS, donc un sous-ensemble de l'ensemble des valeurs ξS obtenues pour toutes les valeurs possibles de ξ. En vertu du théorème de la borne supérieure d'un sous-ensemble, on trouve:

$$S^{(k)} = \max(S^{(1)}, \ldots, S^{(m)}) \leqslant \max_\xi \xi S. \tag{8-53}$$

D'autre part, en substituant dans l'expression pour ξS aux grandeurs $S^{(1)}, \ldots, S^{(m)}$ leur valeur maximale $S^{(k)}$, on a:

$$\xi S = \sum_i \xi^{(i)} S^{(i)} \leqslant S^{(k)} \sum_i \xi^{(i)} = S^{(k)}. \tag{8-54}$$

Cette expression reste vraie pour tout ξ satisfaisant à la relation (8-16), y compris le cas où ξS atteint son maximum. En comparant (8-53) à (8-54), on arrive à la relation (8-51).

Le théorème qui vient d'être démontré nous permet de mettre l'expression de $B_{\Gamma_1}(S)$ sous la forme

$$B_{\Gamma_1}(S) = \max_\xi \xi S = \max(S^{(1)}, \ldots, S^{(m)}). \tag{8-55}$$

De l'expression (8-55) on tire les corollaires suivants.

C o r o l l a i r e 1. La condition $\beta_\Gamma \leqslant B_{\Gamma_1}(S)$ amène:

$$\beta_\Gamma \leqslant \max(S^{(1)}, \ldots, S^{(m)}). \tag{8-56}$$

Tout point $S \in S^*$ possède au moins une coordonnée non inférieure (donc supérieure ou égale) à β_Γ.

C o r o l l a i r e 2. En posant $S = S_0 = (S_0^{(1)}, \ldots, S_0^{(m)})$ on obtient:

$$\beta_\Gamma = B_{\Gamma_1}(S_0) = \max(S_0^{(1)}, \ldots, S_0^{(m)}). \tag{8-57}$$

La valeur supérieure β_Γ du jeu est égale à la coordonnée maximale du point S_0 qui définit la stratégie minimax du deuxième joueur.

c) Théorème de minimax

La possibilité dont dispose chaque joueur de trouver sa meilleure stratégie est basée sur le théorème suivant qui peut être considéré comme la démonstration de l'existence d'une solution pour les jeux finis.

Théorème 8-5. *Tout jeu fini a une valeur et chaque joueur dispose d'au moins une stratégie optimale.*

Prémisses de départ. Soient $G = (X, Y, L)$ un jeu fini et $\Gamma = (E, H, L)$ le centrage de ce jeu. Lors de la démonstration du théorème il sera commode de raisonner en termes d'un S-jeu. Pour cette raison, désignons par $\Gamma_1 = (E, S^*, L_1)$ le S-jeu équivalent.

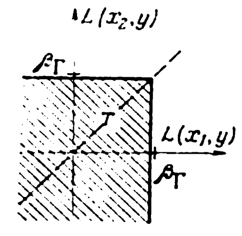


Fig. 8-3. Domaine T dans un espace bidimensionnel

Les valeurs inférieure et supérieure du jeu seront respectivement égales à α_Γ et à β_Γ indépendamment de ce qu'on considère le jeu Γ ou le S-jeu équivalent Γ_1. D'autre part, comme il a été démontré, $\alpha_\Gamma \leqslant \beta_\Gamma$.

Pour démontrer le théorème, il suffit de montrer que $\beta_\Gamma \leqslant \alpha_\Gamma$, car la confrontation avec l'inégalité précédente entraîne $\alpha_\Gamma = \beta_\Gamma$, ce qui signifie que le jeu a une valeur. Mais pour démontrer cette dernière inégalité, il suffit de trouver une stratégie mixte ξ_0 du premier joueur telle que pour tous les $S \in S^*$ la relation ci-dessous ait lieu:

$$\beta_\Gamma \leqslant \xi_0 S, \quad S \in S^*. \tag{8-58}$$

En effet, si (8-58) a lieu, alors

$$\beta_\Gamma \leqslant \min_S \xi_0 S = A_\Gamma(\xi_0) = \alpha_\Gamma. \tag{8-59}$$

Ainsi, la démonstration du théorème sera réduite à la démonstration de l'inégalité (8-58).

Démonstration. Considérons l'ensemble T comportant tous les points $t = (t^{(1)}, \ldots, t^{(m)})$ tels que $t^{(i)} < \beta_\Gamma$ pour $i = 0, 1, \ldots, m$. Sur la figure 8-3 est représenté le domaine T pour un espace bidimensionnel. Dans ce cas, le domaine T a la forme d'un coin rectangulaire dont le sommet se situe sur la droite qui passe par l'origine des coordonnées et qui fait un angle de 45° avec la direction positive de l'axe des abscisses. Elucidons certaines propriétés de l'ensemble T.

L'ensemble T est un ensemble convexe. Considérons deux points arbitraires t_1 et t_2 de cet ensemble. L'équation du segment réunissant ces deux points sera de la forme

$$t = w_1 t_1 + w_2 t_2; \quad w_1, w_2 \geqslant 0; \quad w_1 + w_2 = 1. \tag{8-60}$$

Projetant cette équation sur le i-ème axe et tenant compte du théorème 8-4, on obtient:

$$t^{(i)} = w_1 t_1^{(i)} + w_2 t_2^{(i)} \leqslant \max (t_1^{(i)}, t_2^{(i)}) < \beta_\Gamma. \qquad (8\text{-}61)$$

Il s'ensuit que tout point du segment considéré appartient à l'ensemble T qui est donc un ensemble convexe.

T et S^* sont des ensembles disjoints. Cela découle du fait que tout point de l'ensemble S^* a au moins une coordonnée supérieure ou égale à β_Γ (voir le corollaire 1 du théorème 8-4), donc T et S^* n'ont pas de points communs.

Etant donné que T et S^* sont les domaines convexes disjoints, il y a un hyperplan qui les sépare de façon que les ensembles T et S^* se situent dans des demi-espaces différents définis par cet hyperplan. Il existe donc un $a = (a^{(1)}, \ldots, a^{(m)})$ et un nombre c tels que l'équation

$$ax = c \qquad (8\text{-}62)$$

soit l'équation de cet hyperplan séparateur, avec

$$\left.\begin{aligned} aS \geqslant c \text{ pour } S \in S^*; \\ at \leqslant c \text{ pour } t \in T. \end{aligned}\right\} \qquad (8\text{-}63)$$

Montrons que $a^{(i)} \geqslant 0$, $i = 1, \ldots, m$. Soit $\delta_i = (\delta_i^{(1)}, \ldots, \delta_i^{(m)})$ un point dont la i-ème coordonnée est égale à 1, toutes les autres ayant une petite valeur $\varepsilon > 0$. Considérons le point $S_0 \in S^*$. Etant donné que sa coordonnée maximale est égale à β_Γ (voir le corollaire 2 du théorème 8-4), le point $S_0 - \delta_i \in T$. Donc

$$aS_0 \geqslant c \geqslant a(S_0 - \delta_i). \qquad (8\text{-}64)$$

Il s'ensuit que

$$a\delta_i = a^{(1)}\delta_i^{(1)} + \ldots + a^{(m)}\delta_i^{(m)} \geqslant 0. \qquad (8\text{-}65)$$

Si $\varepsilon \to 0$, alors $\delta_i^{(k)} \to 0$, pour $k \neq i$ et $\delta_i^{(i)} = 1$. Cela étant, la dernière condition donne:

$$a^{(i)} \geqslant 0, \quad i = 1, \ldots, m. \qquad (8\text{-}66)$$

Introduisons la notation

$$\xi_0 = \frac{a}{\sum_i a^{(i)}}. \qquad (8\text{-}67)$$

Il est évident que $\xi_0 \in E$, car

$$\xi_0^{(i)} = \frac{a^{(i)}}{\sum_i a^{(i)}} \geqslant 0, \quad \sum_i \xi_0^{(i)} = 1. \qquad (8\text{-}68)$$

Introduisons en outre la notation

$$\nu = \frac{c}{\sum_i a^{(i)}}. \tag{8-69}$$

Divisons les inégalités (8-63) par $\sum_i a^{(i)}$. Prenant en considération (8-67) et (8-69), on obtient:

$$\left.\begin{aligned} \xi_0 S \geqslant \nu \text{ pour } S \in S^*; \\ \xi_0 t \leqslant \nu \text{ pour } t \in T. \end{aligned}\right\} \tag{8-70}$$

Considérons le point t_0 de coordonnées $t_0^{(i)} = \beta_\Gamma - \varepsilon$, $\varepsilon > 0$, $i = 1, \ldots, m$. Il est évident que $t_0 \in T$. En vertu de la deuxième des inégalités (8-70), on a:

$$\xi_0 t_0 \leqslant \nu. \tag{8-71}$$

Soit $\varepsilon \to 0$, de sorte que $t_0^{(i)} \to \beta_\Gamma$. Alors

$$\xi_0 t_0 = \sum_i \xi_0^{(i)} t_0^{(i)} \to \beta_\Gamma \sum_i \xi_0^{(i)} = \beta_\Gamma. \tag{8-72}$$

En comparant (8-71) à (8-72), on trouve:

$$\nu \geqslant \beta_\Gamma. \tag{8-73}$$

Dans ce cas, la première des inégalités (8-70) donne:

$$\xi_0 S \geqslant \nu \geqslant \beta_\Gamma, \tag{8-74}$$

ce qui démontre l'inégalité (8-58).

De cette façon, $\nu = \alpha_\Gamma = \beta_\Gamma$ est la valeur du jeu, ξ_0 et S_0 représentant les stratégies mixtes optimales des joueurs.

d) Représentation géométrique du principe du minimax

Montrons que le point S_0 qui définit la stratégie minimax du deuxième joueur est un point frontière du domaine S^* et tel qu'en ce point le domaine T touche le domaine S^*.

Comme on l'a vu, $S_0 = (S_0^{(1)}, \ldots, S_0^{(m)}) \in S^*$ avec $\max (S_0^{(1)}, \ldots, S_0^{(m)}) = \beta_\Gamma$. D'autre part, $t = (t^{(1)}, \ldots, t^{(m)}) \in T$ si $t^{(i)} < \beta_\Gamma$, $i = 1, \ldots, m$.

Considérons le point $S_0' = (S_0^{(1)} - \varepsilon, \ldots, S_0^{(m)} - \varepsilon)$ où $\varepsilon > 0$. Il est évident que $\max (S_0^{(1)} - \varepsilon, \ldots, S_0^{(m)} - \varepsilon) < \beta_\Gamma$, de sorte que $S_0^{(i)} - \varepsilon < \beta_\Gamma$, $i = 1, \ldots, m$. Il s'ensuit que $S_0' \in T$. Mais $\lim_{\varepsilon \to 0} S_0' = S_0 \in S^*$.

D'ici on tire deux conclusions:

1) S_0 est un point frontière du domaine S^*;
2) S_0 est un point de contact du domaine T avec le domaine S^*.

Ces propriétés permettent de trouver facilement à l'aide d'une construction géométrique la stratégie minimax S_0 pour le cas où le premier joueur dispose de deux stratégies pures, c.-à-d. lorsque le S-jeu équivalent est représenté par un ensemble de points situé dans le plan. Pour construire le domaine T tangent au domaine S^*, il est commode de tracer sous un angle de 45° par rapport à l'axe des abscisses une droite passant par l'origine des coordonnées et portant le sommet du coin rectangle qui forme le domaine T.

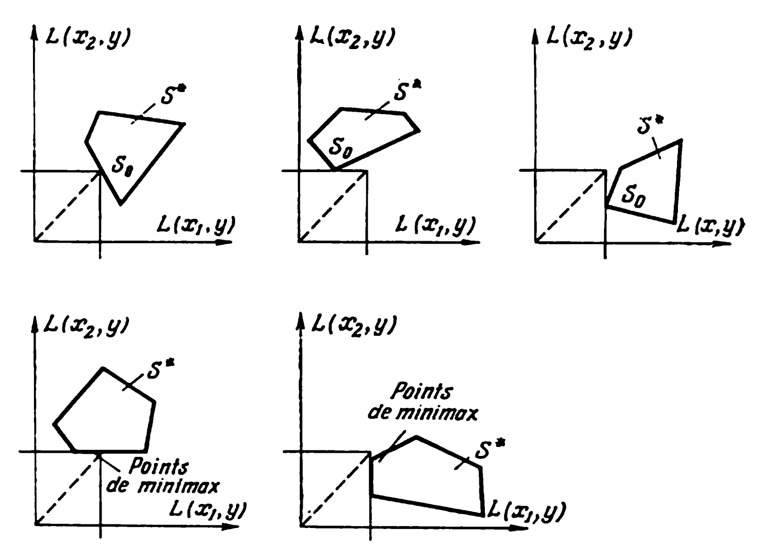


Fig. 8-4. Détermination géométrique de la stratégie minimax

Sur la figure 8-4 sont représentés divers cas de disposition réciproque des domaines S^* et T et sont marqués les points qui définissent la stratégie minimax S_0 du deuxième joueur.

8-4. SOLUTION DES JEUX

a) Stratégies dominantes et stratégies utiles

On sait que les stratégies mixtes des joueurs représentent un mélange des stratégies pures $x \in X$ et $y \in Y$ prises conformément à la distribution de probabilités $\xi(x)$ et $\eta(y)$. Mais dans beaucoup de cas, il est de toute évidence que l'adoption de certaines stratégies pures n'est pas rationnelle et, lors de la détermination de la stratégie mixte optimale, ces stratégies doivent tout simplement être écartées. Nous allons appeler *stratégies utiles* du joueur les stratégies pures qui contribuent à l'élaboration de sa stratégie mixte

optimale. Pour faciliter la séparation des stratégies utiles, introduisons la notion de stratégie dominante.

Considérons deux stratégies y_l et y_p du deuxième joueur. Supposons que le premier joueur adopte la stratégie x_i. Les pertes du deuxième joueur seront respectivement égales à q_{il} et q_{ip}. Il se peut que pour n'importe quel i

$$q_{il} \leqslant q_{ip}, \quad i = 1, \ldots, m, \tag{8-75}$$

c.-à-d. que dans la matrice du jeu les pertes figurant dans la colonne l ne dépassent pas les pertes correspondantes de la colonne p. Cela

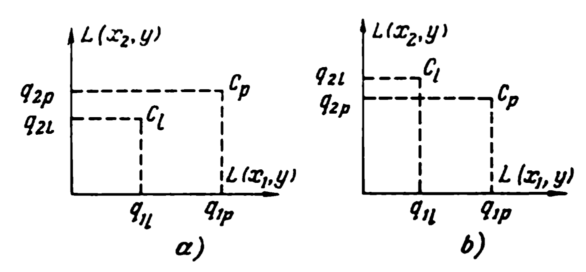


Fig. 8-5. Détermination de la dominance dans un S-jeu

signifie que le deuxième joueur ne doit en aucun cas adopter la stratégie y_p, car en l'adoptant il subira à coup sûr des pertes supérieures à celles qu'entraînerait la stratégie y_l. C'est pourquoi la stratégie y_p doit être écartée, c.-à-d. rayée de la matrice du jeu. La stratégie y_l qui satisfait à la condition (8-75) est appelée stratégie *dominante* par rapport à la stratégie y_p.

Les stratégies dominantes du deuxième joueur sont bien mises en évidence par une construction géométrique en passant au S-jeu équivalent dans le plan. Dans ce cas, $m = 2$ et la condition (8-75) s'écrira

$$q_{1l} \leqslant q_{1p}, \quad q_{2l} \leqslant q_{2p}. \tag{8-76}$$

Sur la figure 8-5 sont donnés deux cas de disposition des points C_l et C_p correspondant aux stratégies pures y_l et y_p du deuxième joueur. On voit sans difficulté que, dans le cas représenté par la figure 8-5, *a*, la stratégie C_l domine la stratégie C_p, tandis que dans le cas de la figure 8-5, *b*, aucune des stratégies n'est une stratégie dominante. Pour que la stratégie y_l domine la stratégie y_p, le point C_l doit se situer à gauche et plus bas par rapport au point C_p.

D'une façon analogue sont déterminées les stratégies dominantes du premier joueur. On dit que la stratégie x_l domine la stratégie x_p si le gain du premier joueur, qui adopte la stratégie x_l, est supé-

rieur aux gains apportés par l'utilisation de la stratégie x_p, quelle que soit la stratégie $y \in Y$:

$$q_{li} \geqslant q_{pi}, \quad i=1, \ldots, n, \tag{8-77}$$

c.-à-d. si dans la matrice du jeu les pertes figurant dans la ligne x_l sont supérieures aux pertes correspondantes de la ligne x_p.

Exemple 8-3. Sur la figure 8-6, un S-jeu est donné dans le plan par les points C_1 à C_6. La figure montre que le point C_1 domine le point C_5, tandis que les points C_1, C_2 et C_6 dominent le point C_4. En écartant les points C_5 et C_4, on arrive à un S-jeu défini par les points C_1, C_2, C_3, C_6 parmi lesquels il n'y a pas de points dominants.

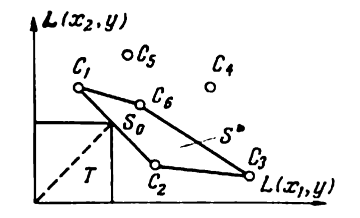


Fig. 8-6. Exemple d'un S-jeu

En éliminant de la matrice du jeu les stratégies dominées par d'autres stratégies, on simplifie considérablement le jeu, donc la recherche de la stratégie optimale. Supposons maintenant que dans le jeu considéré il n'y ait pas de stratégies dominantes. Dans ce cas, on peut se demander si toutes les stratégies sont utiles, c.-à-d. si elles sont utilisées toutes pour l'obtention de la stratégie mixte optimale.

Adressons-nous au jeu représenté sur la figure 8-6. Après avoir éliminé les points C_4 et C_5 dominés par d'autres points, on est arrivé au jeu représenté par les points C_1, C_2, C_3, C_6. Ayant construit le domaine T, on voit que le point S_0, qui définit la stratégie optimale du deuxième joueur, se situe sur la droite réunissant les points C_1 et C_2 et peut être présenté comme la moyenne pondérée de ces deux points. De cette façon, la stratégie mixte optimale du deuxième joueur représente un mélange des stratégies pures y_1 et y_2 qui, dans ce cas, sont des stratégies utiles. Les stratégies y_3 et y_6 ne sont pas des stratégies utiles et leur utilisation n'est pas rationnelle.

Le nombre total de stratégies utiles de chaque joueur peut être déterminé en partant du fait que l'enveloppe convexe S^* d'un ensemble fini $\{C_1, \ldots, C_n\}$ est un polyèdre convexe dans un espace à m dimensions. Le point S_0, qui est un point frontière du polyèdre convexe S^*, appartiendra obligatoirement à une de ses faces dont les sommets correspondront aux stratégies utiles du deuxième joueur. Ayant en vue que le nombre de sommets de n'importe quelle face du polyèdre convexe S^* ne peut pas être supérieur au nombre total de ses sommets, c.-à-d. à n, et ne peut pas dépasser la dimension de l'espace dans lequel se situe ce polyèdre convexe, c.-à-d. m, on arrive à la conclusion que le nombre de stratégies utiles du deuxième joueur n'est pas supérieur au plus petit des nombres m et n. Etant donné que les notions de premier et de deuxième joueur ne sont que conventionnelles, une conclusion analogue reste vraie également pour le premier joueur.

De cette façon, dans un jeu à matrice de dimension $m \times n$ le nombre de stratégies utiles de chacun des joueurs n'est pas supérieur au plus petit des nombres m et n.

Théorème 8-6. *Si un des joueurs suit en permanence sa stratégie mixte optimale, le gain des joueurs reste invariable et égal à la valeur du jeu* v *indépendamment de la stratégie mixte (ou pure) adoptée par l'autre joueur si seulement ce dernier ne sort pas du domaine comportant ses stratégies utiles.*

Démonstration. Soient v la valeur du jeu et $\xi_0 = (\xi_0^{(1)}, \ldots, \xi_0^{(m)})$ la stratégie mixte optimale du premier joueur. Si ce dernier adopte sa stratégie optimale ξ_0, son gain ne peut pas être inférieur à v.

Supposons que le deuxième joueur dispose de k stratégies utiles désignées par $y_1, \ldots, y_k$. Si le premier joueur adopte sa stratégie optimale ξ_0, tandis que le deuxième utilise sa stratégie utile y_i $(i = 1, \ldots, k)$, alors le gain du premier joueur v_i ne sera pas inférieur à v:

$$v_i \geqslant v, \quad i = 1, \ldots, k. \tag{8-78}$$

Examinons maintenant la stratégie mixte optimale η_0 du deuxième joueur. Elle est donnée par les probabilités $\eta_0^{(i)}$ avec lesquelles sont utilisées les stratégies utiles pures. Il est essentiel que l'on ne peut pas avoir $\eta_0^{(i)} = 0$ pour $i = 1, \ldots, k$, car, dans ce cas, la i-ème stratégie ne contribuerait pas à la formation de la stratégie mixte optimale du deuxième joueur, donc ne serait pas une stratégie utile.

Trouvons le gain moyen du premier joueur dans le cas où le deuxième adopte sa stratégie mixte optimale. Lorsque le deuxième joueur adopte sa stratégie mixte optimale η_0, cela signifie que les gains $v_1, \ldots, v_k$ obtenus par le premier joueur correspondront aux stratégies pures $y_1, \ldots, y_k$ utilisées avec les probabilités $\eta_0^{(i)}, \ldots, \eta_0^{(k)}$. Le gain moyen du premier joueur sera donc égal à

$$\underline{L}_m = v_1\eta_0^{(1)} + \ldots + v_k\eta_0^{(k)} = \sum_{i=1}^{k} v_i\eta_0^{(i)}. \tag{8-79}$$

En substituant v à v_i et compte tenu de (8-78), on obtient:

$$L_m \leqslant v \sum_{i=1}^{k} \eta_0^{(i)} = v. \tag{8-80}$$

Mais le gain moyen obtenu lors de l'utilisation des stratégies optimales n'est autre chose que la valeur du jeu v, c.-à-d.

$$L_m = v. \tag{8-81}$$

D'autre part, la relation (8-79) ne peut se transformer en (8-81) que si la condition

$$v_i = v, \; i = 1, \; \ldots, \; k, \qquad (8\text{-}82)$$

est satisfaite, ce qui démontre le théorème.

On voit de cette façon que le gain moyen du premier joueur, qui adopte sa stratégie mixte optimale, restera le même quelle, que soit la stratégie utile, et donc toute stratégie mixte composée de stratégies utiles, adoptée par le deuxième joueur.

Il ne faut pas oublier qu'en toute circonstance l'utilisation de la stratégie mixte optimale est la plus avantageuse. L'adoption des stratégies utiles, même non optimales, ne cause pas de préjudices seulement dans le cas où l'adversaire suit sa stratégie optimale, tandis que l'utilisation de la stratégie optimale assure toujours un gain égal à la valeur du jeu.

Il nous reste maintenant à examiner les procédés qui permettent de trouver les stratégies utiles et optimales des joueurs, c.-à-d. d'aboutir à la solution du jeu.

b) Recherche des stratégies optimales

Considérons un jeu $G = (X, Y, L)$ à matrice $m \times n$ qui n'a pas de point-selle (dans le cas contraire, il est facile de trouver la solution du jeu selon la règle établie au paragraphe 8-2) et duquel on a éliminé les stratégies dominées par d'autres stratégies.

Désignons par $\xi_0 = (\xi_0^{(1)}, \ldots, \xi_0^{(m)})$ la stratégie mixte optimale du premier joueur. Certains $\xi_0^{(i)}$ peuvent être nuls, ce qui signifie que la stratégie correspondante x_i n'est pas utile. Il nous faut trouver les valeurs de $\xi_0^{(i)}$ pour $i = 1, \ldots, m$.

Supposons que le deuxième joueur utilise la stratégie y_k. Si c'est une stratégie utile, le gain du premier joueur sera v, dans le cas contraire, il pourra dépasser cette valeur. Donc, dans le cas général, on a:

$$L(\xi_0, y_k) = \xi_0^{(1)} q_{1k} + \ldots + \xi_0^{(m)} q_{mk} \geqslant v. \qquad (8\text{-}83)$$

Des expressions pareilles peuvent être écrites pour chaque stratégie pure du deuxième joueur, c.-à-d. pour $i = 1, \ldots, n$. On sait en outre que

$$\xi_0^{(i)} \geqslant 0, \quad \xi_0^{(1)} + \ldots + \xi_0^{(m)} = 1. \qquad (8\text{-}84)$$

Posons $v > 0$. Cela sera toujours vrai si tous les éléments q_{ik} de la matrice du jeu sont positifs. Si certains des éléments q_{ik} sont négatifs, on peut les rendre positifs en ajoutant à tous les éléments de la matrice un certain nombre $v' > 0$. Dans ce cas, la valeur du jeu augmentera aussi de la quantité v'.

Divisons les équations (8-83) et (8-84) par ν en notant

$$\frac{\xi_0^{(i)}}{\nu}=p_i. \tag{8-85}$$

Il est évident que $p_i \geqslant 0$. Dans ce cas, les inégalités (8-83) s'écriront, pour $k = 1, \ldots, n$, sous la forme

$$\left.\begin{array}{l} p_1q_{11}+\ldots+p_mq_{m1}\geqslant 1; \\ \ldots\ldots\ldots\ldots\ldots \\ p_1q_{1n}+\ldots+p_mq_{mn}\geqslant 1, \end{array}\right\} \tag{8-86}$$

tandis que la relation (8-84) deviendra

$$p_1+\ldots+p_m=\frac{1}{\nu}. \tag{8-87}$$

Les inégalités linéaires (8-86) contiennent les inconnues $p_1, \ldots, p_m$ qui définissent la stratégie mixte du premier joueur. La détermination de la stratégie mixte optimale implique une telle solution du système (8-86) pour laquelle la quantité ν devient maximale et la forme linéaire (8-87) atteint donc son minimum. Mais c'est un problème ordinaire de programmation linéaire.

Dans les expressions (8-86) il est plus commode de passer des inégalités aux égalités en soustrayant des seconds membres les inconnues additionnelles positives z_i, ce qui aboutit au système d'équations

$$p_1q_{1i}+\ldots+p_mq_{mi}-z_i=1, \quad i=1, \ldots, n. \tag{8-88}$$

En résolvant les équations (8-88) correspondant aux stratégies utiles du deuxième joueur on obtient $z_i = 0$. De cette façon, en résolvant le système (8-88) sous la condition de minimisation de la forme linéaire (8-87), on trouve la stratégie mixte du premier joueur, la valeur du jeu ν et les stratégies utiles $y_1, \ldots, y_k$ du deuxième joueur.

Pour trouver la stratégie mixte optimale du deuxième joueur $\eta_0 = (\eta_0^{(1)}, \ldots, \eta_0^{(k)})$ comportant k inconnues $\eta_0^{(1)}, \ldots, \eta_0^{(k)}$, il est nécessaire de former k équations. Une de ces équations est

$$\eta_0^{(1)}+\ldots+\eta_0^{(k)}=1. \tag{8-89}$$

Les autres $k - 1$ équations seront obtenues en formant les expressions des pertes moyennes pour la stratégie mixte optimale du deuxième joueur et pour n'importe quelles $k-1$ stratégies utiles du premier joueur. Ainsi, pour la stratégie utile x_i on aura l'équation

$$\eta_0^{(1)}q_{i1}+\ldots+\eta_0^{(k)}q_{ik}=\nu. \tag{8-90}$$

En formant k — 1 équations de la forme (8-90) et en leur ajoutant l'équation (8-89), on obtient k équations dont la résolution permet de trouver sans difficulté les quantités $\eta_0^{(1)}, \ldots, \eta_0^{(k)}$ qui déterminent la stratégie mixte optimale η_0 du deuxième joueur.

Exemple 8-4. Trouver la solution du jeu dont la matrice est donnée par le tableau 8-6.

Avant tout, il faut se convaincre que le jeu n'a pas de point-selle et qu'il n'y a pas de stratégies qui dominent autres stratégies. Pour éliminer les éléments négatifs de la matrice du jeu, ajoutons à chaque élément de la matrice e nombre 5. Alors, la matrice du jeu prendra la forme du tableau 8-7.

Tableau 8-6

	y_1	y_2	y_3
x_1	2	—3	4
x_2	—3	4	—5
x_3	4	—5	6

Tableau 8-7

	y_1	y_2	y_3
x_1	7	2	9
x_2	2	9	0
x_3	9	0	11

Les équations (8-88) s'écriront sous la forme

$$7p_1 + 2p_2 + 9p_3 - z_1 = 1;$$
$$2p_1 + 9p_2 - z_2 = 1;$$
$$9p_1 + 11p_3 - z_3 = 1.$$

La solution de ces équations doit satisfaire à la condition de minimum de la forme linéaire (8-87)

$$p_1 + p_2 + p_3 = \min.$$

En résolvant le problème de programmation linéaire obtenu, on trouve:

$$z_1 = z_2 = z_3 = 0; \; p_1 = 0{,}05; \; p_2 = 0{,}1; \; p_3 = 0{,}05.$$

On voit donc que toutes les trois stratégies y_1, y_2 et y_3 sont des stratégies utiles. De l'expression (8-87) on tire:

$$v = \frac{1}{p_1 + p_2 + p_3} = 5.$$

Dans ce cas,

$$\xi_0^{(1)} = 5p_1 = 0{,}25; \quad \xi_0^{(2)} = 5p_2 = 0{,}5; \quad \xi_0^{(3)} = 5p_3 = 0{,}25.$$

Pour trouver la stratégie mixte optimale du deuxième joueur, on compose une équation de la forme (8-89)

$$\eta_0^{(1)} + \eta_0^{(2)} + \eta_0^{(3)} = 1$$

et deux équations de la forme (8-90) pour les stratégies utiles du premier joueur x_2 et x_3

$$\left.\begin{aligned} 2\eta_0^{(1)} + 9\eta_0^{(2)} &= 5; \\ 9\eta_0^{(1)} + 11\eta_0^{(3)} &= 5. \end{aligned}\right\}$$

En résolvant le système comportant les trois équations obtenues, on trouve :

$$\eta_0^{(1)} = \eta_0^{(3)} = 0{,}25; \quad \eta_0^{(2)} = 0{,}5.$$

Compte tenu du nombre 5 qui a été ajouté à chaque élément de la matrice, on trouve la valeur du jeu $v - 5 = 0$.

c) Représentation géométrique du principe du minimax dans le jeu $2 \times n$

Dans le jeu $2 \times n$, la stratégie mixte du premier joueur représente le couple ordonné $\xi = (\zeta, 1 - \zeta)$. Si le deuxième joueur adopte la stratégie y_k, alors le gain moyen du premier joueur sera

Tableau 8-8

	y_1	y_2	y_3
x_1	1	3	5
x_2	4	2	1

$$L(\zeta, y_k) = \zeta q_{1k} + (1 - \zeta) q_{2k} = \\ = q_{2k} + \zeta (q_{1k} - q_{2k}), \quad (8\text{-}91)$$

c.-à-d. il dépend linéairement de la grandeur ζ, comme le montre la figure 8-7.

Examinons la détermination de la stratégie optimale ξ_0 à l'aide du jeu dont la matrice est donnée par le tableau 8-8. Sur la figure 8-8 les chiffres *1*, *2* et *3* désignent respectivement les lignes de gains pour les stratégies y_1, y_2 et y_3. Le gain garanti du premier

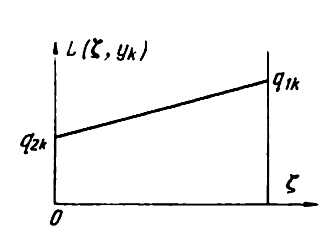


Fig. 8-7. Ligne de gains pour la stratégie y_k

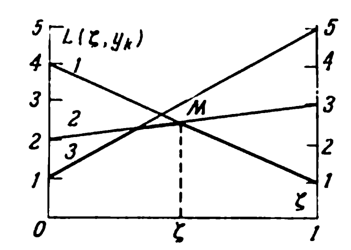


Fig. 8-8. Stratégie minimax dans le jeu $2 \times n$

joueur pour tout ζ sera déterminé par la borne inférieure (en trait fort) du graphique donné. Conformément au principe du minimax, la stratégie optimale ζ_0 doit assurer le gain garanti maximal. Cette stratégie est déterminée par le point *M* situé à l'intersection des lignes de gains correspondant aux stratégies y_1 et y_2 qui, de cette façon, seront les stratégies utiles du deuxième joueur.

La stratégie mixte optimale ξ_0 peut être trouvée en partant de la condition d'égalité du gain moyen du premier joueur lors de l'utilisation des stratégies utiles y_1 et y_2 du deuxième joueur

$$\zeta_0 + 4(1 - \zeta_0) = 3\zeta_0 + 2(1 - \zeta_0),$$

ce qui donne $\zeta_0 = 1/2$. Ainsi, $\xi_0 = (1/2, 1/2)$. Maintenant, la valeur du jeu sera trouvée en tant que gain moyen du premier joueur utilisant la stratégie ξ_0 pour n'importe quelle stratégie utile, par exemple y_1, du deuxième joueur

$$\nu = \zeta_0 + 4(1 - \zeta_0) = 2,5.$$

Si la valeur du jeu ν est connue, la stratégie mixte optimale $\eta_0 = (\eta_0^{(1)}, 1 - \eta_0^{(1)}, 0)$ du deuxième joueur sera trouvée en partant de l'expression des pertes moyennes du deuxième joueur pour n'importe quelle stratégie utile du premier joueur. Par exemple, pour la stratégie x_1, on a:

$$\eta_0^{(1)} + 3(1 - \eta_0^{(1)}) = 2,5,$$

ce qui donne $\eta_0^{(1)} = 1/4$. Donc, $\eta_0 = (1/4, 3/4, 0)$.

PROBLÈMES AU CHAPITRE 8

8-1. Calculer la matrice de paiement pour le jeu de l'exemple 8-1.

8-2. Calculer la matrice de paiement pour le jeu ci-après ayant dessiné au préalable l'arbre de ce jeu.

On jette une pièce de monnaie symétrique et si le jet amène pile, le premier joueur choisit un des nombres 1, 6. Si le jet amène face, ce joueur choisit un des nombres 2, 7. Ensuite, le deuxième joueur choisit un des deux nombres 3, 9. Les nombres choisis par les joueurs sont additionnés et donnent la somme S. Ensuite, on tire au sort avec la probabilité d'amener pile de 0,8. Si le sort amène pile, le deuxième joueur paie au premier S roubles, si c'est face, c'est le premier joueur qui paie au deuxième S roubles. Le premier joueur connaît l'issue du premier coup aléatoire. Le deuxième joueur ne sait rien sur les coups précédents.

8-3. Un mécanisme aléatoire commode à utiliser est offert par la position qu'occupe la trotteuse sur le cadran d'une montre à un instant donné choisi au hasard. Décrire l'utilisation de ce mécanisme aléatoire pour obtenir les distributions de probabilités suivantes:

$$\left(\frac{3}{5}, \frac{2}{5}\right), \quad \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right), \quad \left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right).$$

8-4. Décrire les actions des joueurs participant à une partie du jeu dont la matrice est donnée par le tableau 8-4 en sachant que le premier joueur adopte la stratégie pure x_2 tandis que le deuxième utilise la stratégie mixte $\eta = (0,5, 0,5)$.

CHAPITRE 9

THÉORIE DES DÉCISIONS STATISTIQUES (JEUX STATISTIQUES)

9-1. STRUCTURE DES JEUX STATISTIQUES

a) Jeux stratégiques et jeux statistiques

Un type spécifique de jeux particulièrement importants lors de l'analyse des différentes situations pratiques est représenté par les *jeux* dits *statistiques*. Ces jeux diffèrent essentiellement du type de jeux examinés jusqu'à présent qui pourraient être appelés *jeux stratégiques*.

A la base de la théorie des jeux stratégiques on trouve l'hypothèse des intérêts diamétralement opposés de deux joueurs. Chacun des joueurs tâche de choisir sa stratégie de façon à s'assurer le profit maximal tout en réduisant au minimum le profit de l'adversaire. Dans de pareils jeux, chaque joueur agit activement et tend autant que possible à utiliser sa stratégie optimale.

Pourtant, dans beaucoup de situations pratiques on rencontre des cas où l'un des joueurs est *neutre*, c.-à-d. qu'il ne tend pas à tirer le profit maximal, donc ne cherche pas à mettre à profit les erreurs commises par son adversaire. A ce type de jeux se rapportent les jeux dont un des participants est représenté par la nature. Ici, le mot « nature » signifie tout l'ensemble des circonstances extérieures qui déterminent la prise des décisions.

La nature ne peut pas être considérée en tant qu'adversaire raisonnable capable de profiter des erreurs commises par l'homme. Autrement dit, la nature est sans mauvaise intention à l'égard de l'homme. Elle évolue et agit conformément à ses lois et c'est l'homme qui a la possibilité de mettre à son profit ces lois. Si l'homme connaissait parfaitement les lois de la nature, il pourrait en tirer l'avantage maximal. Mais dans beaucoup de cas, l'homme ne connaît pas les lois de la nature ou les connaît insuffisamment.

Le prix qu'il faut inévitablement payer en tâchant d'obtenir une décision, sans avoir une information complète sur la loi de la nature, consiste dans la possibilité de prendre des décisions erronées. D'autre part, les situations pratiques sont parfois telles qu'il est impossible de renoncer à la prise de décision. En outre, la décision de renoncer à la prise de décision est aussi une décision qui peut entraîner des conséquences non moins indésirables que celles entraînées par les autres décisions. L'unique possibilité de se tirer d'affaire dans cette situation réside en l'élaboration par l'homme d'une telle

stratégie de la prise de décision qui, tout en admettant l'éventualité d'une prise de décisions erronées, réduise au minimum les conséquences indésirables qui en découlent.

Il est vrai que l'homme peut encore étudier son adversaire (la nature) en faisant des expériences.

Théoriquement, en faisant des expériences non limitées, on peut compléter autant que l'on veut ses connaissances de la nature pour agir ensuite en toute netteté, mais deux circonstances s'y opposent :

les expériences nécessitent du temps, tandis que dans beaucoup de cas la décision doit être prise rapidement ;

les expériences impliquent des frais, donc peuvent coûter plus cher que l'avantage apporté par le surplus de connaissances qu'elles fournissent.

C'est pourquoi, dans le jeu de l'homme contre la « nature », un problème important est constitué par la prise de décision concernant les expériences : faut-il les faire, et, s'il le faut, quelles sont les expériences qu'il faut faire, quand doit-on les arrêter et quelles sont les actions à entreprendre une fois les expériences finies ?

Les jeux où l'un des adversaires est la nature, l'autre étant l'homme, ont été appelés *jeux statistiques*, leur théorie étant dite *théorie des décisions statistiques* [54, 57, 58]. L'homme qui participe au jeu sera appelé, dans ce qui suit, *statisticien*.

b) Espace des stratégies de la nature

Par stratégie de la nature nous allons entendre l'ensemble complet des conditions extérieures présentes à la prise d'une décision. Cet ensemble des conditions extérieures sera appelé état de la nature ϑ. Dans le cas général, il existe un certain ensemble des états possibles de la nature $\Theta = \{\vartheta_1, \ldots, \vartheta_m\}$ qui, comme il a été convenu au chapitre 6, s'appellera espace des états de la nature. Les éléments ϑ_i de cet espace seront les stratégies pures de la nature.

Si nous savions au préalable la stratégie pure que la nature va adopter dans chaque cas concret, nous prendrions la décision avec certitude ayant une connaissance parfaite de l'état de la nature. Mais d'habitude, on ne connaît que la collection de stratégies pures de la nature. En outre, l'expérience passée nous apprend la fréquence à laquelle la nature adopte telle ou telle stratégie pure dont elle dispose, c.-à-d. on connaît la distribution a priori de probabilités $\xi(\vartheta)$ sur l'espace des états de la nature Θ. Cette distribution a priori de probabilités $\xi(\vartheta)$ sera appelée *stratégie mixte de la nature*.

c) Espace des stratégies du statisticien et fonction de pertes

Le problème que le statisticien doit résoudre consiste dans la prise d'une certaine décision ou dans l'exécution d'une certaine action appartenant à l'ensemble des décisions ou des actions.

Désignons les actions possibles du statisticien par $a_1, \ldots, a_l$. Chacune de ces actions est une *stratégie pure du statisticien*. L'ensemble $A = \{a_1, \ldots, a_l\}$ est l'*espace des stratégies pures du statisticien*.

Le statisticien doit pouvoir estimer chacune de ses actions. Pour cela, il admet qu'en accomplissant l'action a, il peut encourir la perte $L(\vartheta, a)$ qui est fonction aussi bien de l'action exécutée a que de l'état de la nature ϑ que le statisticien ignore. La fonction $L(\vartheta, a)$, appelée *fonction de pertes*, doit être définie à l'avance pour toutes les combinaisons $a \in A$ et $\vartheta \in \Theta$ possibles, c.-à-d. doit être donnée sur le produit direct des ensembles $\Theta \times A$. Elle peut être donnée analytiquement, ou bien, par analogie à (8-7), à l'aide d'une matrice de paiement de la forme

$$\mathbf{Q} = \| q_{ij} \| = \begin{Vmatrix} q_{11} & \ldots & q_{1l} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ q_{m1} & \ldots & q_{ml} \end{Vmatrix}, \qquad (9\text{-}1)$$

où $q_{ij} = L(\vartheta_i, a_j)$. La connaissance de la fonction de pertes permet au statisticien de prendre les mesures les plus indiquées compte tenu de l'information qu'il possède sur l'état de la nature.

D'habitude, le statisticien connaît la stratégie mixte de la nature, c.-à-d. la distribution a priori de probabilités $\xi(\vartheta)$ sur l'espace des états de la nature Θ. La connaissance de la distribution a priori de probabilités donne la possibilité de déterminer les pertes moyennes subies par le statisticien qui entreprend telle ou telle action:

$$L(\xi, a) = M[L(\vartheta, a)] = \sum_{\vartheta \in \Theta} L(\vartheta, a)\, \xi(\vartheta). \qquad (9\text{-}2)$$

L'action la plus favorable pour le statisticien sera l'action dite *de Bayes* a^* pour laquelle les pertes seront minimales et égales à

$$R^*(\xi) = L(\xi, a^*) = \min_{a \in A} L(\xi, a). \qquad (9\text{-}3)$$

Il n'est pas obligatoire que le statisticien se borne à une seule stratégie pure. Il peut aussi utiliser un mélange de stratégies pures conformément à une certaine loi de distribution de probabilités. Dans ce cas, il sera question de la *stratégie mixte du statisticien*. Pour utiliser une stratégie mixte, le statisticien doit se donner la distribution de probabilités $\eta(a) = (\eta^{(1)}, \ldots, \eta^{(m)})$ qui définit les probabilités avec lesquelles il doit utiliser ses stratégies pures $a_1, \ldots$

. . ., a_m. Dans le cas général, le statisticien dispose d'une certaine collection de stratégies mixtes $H = \{\eta_1(a), \ldots, \eta_\nu(a)\}$ appelée *espace des stratégies mixtes du statisticien.*

Si le statisticien adopte la stratégie mixte $\eta(a)$, tandis que la nature utilise la stratégie mixte $\xi(\vartheta)$, les pertes moyennes du statisticien seront

$$L(\xi, \eta) = M_{\vartheta, a}[L(\vartheta, a)] = \sum_{\vartheta, a} L(\vartheta, a)\,\xi(\vartheta)\,\eta(a). \qquad (9\text{-}4)$$

Dans ce cas, la tâche que s'impose le statisticien consiste à choisir une stratégie $\eta^*(a) \in H$ telle que ses pertes moyennes $L(\xi, \eta^*)$ soient minimales, c.-à-d.

$$L(\xi, \eta^*) = \min_{\eta \in H} L(\xi, \eta). \qquad (9\text{-}5)$$

Les cas qui viennent d'être exposés impliquent la résolution d'un problème statistique relativement simple consistant dans la détermination de la meilleure stratégie du statisticien sur la base seulement de l'information a priori qu'il possède sur les états de la nature. Ici, le statisticien ne tâche pas de préciser l'état réel de la nature en faisant des expériences. C'est pourquoi ce type de jeu statistique peut être appelé *jeu statistique sans expérience.*

d) Exemples de jeux statistiques

Pour mieux comprendre la structure des jeux statistiques et les méthodes de résolution utilisées, nous allons donner quelques exemples qui nous serviront en même temps d'illustration aux points fondamentaux de la théorie des décisions statistiques.

Exemple 9-1. Problème du remplacement de l'équipement. L'équipement complexe et coûteux d'une entreprise ayant servi pendant k ans peut se trouver dans l'un des trois états suivants:

ϑ_1 : l'équipement peut rester en service et ne nécessite qu'une petite réparation courante;

ϑ_2 : certaines pièces sont considérablement usées et nécessitent une réparation importante ou bien doivent être remplacées;

ϑ_3 : les pièces principales sont tellement usées que l'utilisation ultérieure de l'équipement est impossible.

L'expérience passée acquise dans le domaine d'exploitation d'un équipement analogue montre que l'état ϑ_1 caractérise 20 % des cas, l'état ϑ_2 50 % et l'état ϑ_3 30 %.

L'entreprise peut choisir l'une des trois voies suivantes:

a_1 : utiliser l'ancien équipement pendant une année en effectuant une réparation peu importante par ses propres moyens;

a_2 : effectuer une grosse réparation de l'équipement à l'aide d'une équipe spécialisée;

a_3 : remplacer l'équipement ancien par un équipement neuf.

Les pertes subies par l'entreprise et correspondant aux différentes voies sont données au tableau 9-1. Ces pertes incluent les frais de réparation ou de remplacement de l'équipement de même que les dommages liés à la baisse de

la qualité de la production et au temps mort dû aux pannes de l'équipement. Au même tableau on trouve les probabilités a priori des différents états de la nature, autrement dit, la stratégie mixte $\xi(\vartheta)$ de la nature.

Tableau 9-1

Probabilités a priori des états de la nature et pertes dans le problème du remplacement de l'équipement

ϑ	$\xi(\vartheta)$	A		
		a_1	a_2	a_3
ϑ_1	0,2	1	3	5
ϑ_2	0,5	5	2	4
ϑ_3	0,3	7	6	3

Tableau 9-2

Probabilités a priori des états de la nature et pertes dans le problème de la ligne technologique

ϑ	$\xi(\vartheta)$	A		
		a_1	a_2	a_3
ϑ_1	0,6	0	1	3
ϑ_2	0,4	5	3	2

Pour la stratégie mixte donnée $\xi(\vartheta)$, les pertes moyennes correspondant aux différentes voies adoptées sont:

$$L(\xi, a_1) = \sum_{\vartheta} L(\vartheta, a_1)\,\xi(\vartheta) = 1\cdot 0{,}2 + 5\cdot 0{,}5 + 7\cdot 0{,}3 = 4{,}8;$$

$$L(\xi, a_2) = 3{,}4; \quad L(\xi, a_3) = 3{,}9.$$

Exemple 9-2. Problème de la ligne technologique. Une ligne technologique peut recevoir de la matière première contenant peu (ϑ_1) ou beaucoup (ϑ_2) d'impuretés. On sait qu'en moyenne la ligne reçoit 60 % de matière première du premier type et 40 % du deuxième. Pour utiliser des différents types de matière première, la ligne technologique peut fonctionner en un des trois régimes a_1, a_2 et a_3. Les probabilités a priori des états de la nature et les pertes reflétant la qualité du produit fabriqué et la consommation de matière première en fonction de la qualité de cette dernière et du régime de fonctionnement de la ligne technologique sont données au tableau 9-2.

Les pertes moyennes correspondant aux probabilités a priori données pour les différents régimes de fonctionnement sont:

$$L(\xi, a_1) = 2{,}0; \; L(\xi, a_2) = 1{,}8; \; L(\xi, a_3) = 2{,}6.$$

L'action de Bayes correspondra au régime de fonctionnement a_2.

9-2. JEUX STATISTIQUES SANS EXPÉRIENCE

a) Représentation d'un jeu statistique sans expérience sous la forme d'un *S*-jeu

Un jeu statistique peut être représenté sous la forme d'un S-jeu équivalent d'une façon absolument analogue à celle examinée dans le cas des jeux stratégiques. Pour ce faire, à chaque stratégie pure a_j ($j = 1, \ldots, l$) on associe un point $C_j = (q_{1j}, \ldots, q_{mj})$ dans un espace à m dimensions dont les coordonnées seront représentées par les pertes du statisticien $L(\vartheta_i, a_j) = q_{ij}$ pour les différents états

de la nature ϑ_i ($i = 1, \ldots, m$). Ainsi, dans le cas du problème de la ligne technologique, aux stratégies pures du statisticien a_1, a_2 et a_3 seront associés les points $C_1=(0,5)$, $C_2=(1,3)$ et $C_3=(3,2)$ du plan, comme on le voit sur la figure 9-1. L'enveloppe convexe S^* de l'ensemble des points $\{C_1, C_2, C_3\}$ définit le domaine de toutes les stratégies possibles (pures et mixtes) du statisticien.

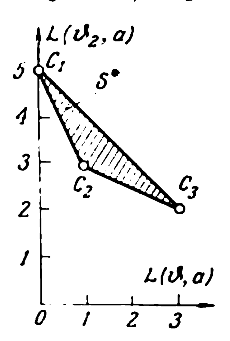


Fig. 9-1. Représentation du problème de la ligne technologique sous la forme d'un S-jeu

Dans ce qui suit, nous allons examiner certains principes qui peuvent guider le statisticien lors du choix de sa stratégie. Il faut quand même remarquer que les statisticiens ne sont pas unanimes sur le meilleur de ces principes lorsqu'il s'agit des jeux statistiques. Autrement dit, il n'y a pas de règle universelle permettant de choisir une certaine façon d'agir indépendamment de la situation qui se présente. Toutefois, quoiqu'il puisse y avoir des divergences sur les actions à entreprendre dans une situation donnée, on peut arriver à un accord complet sur les actions qui sont à bannir. Pour cela, il faut introduire la notion de stratégie admissible analogue à la notion de stratégie dominante dans les jeux stratégiques.

b) Stratégies admissibles dans les jeux statistiques

Supposons que nous soyons en train d'examiner la stratégie mixte η (a) du statisticien. Deux cas peuvent se présenter:

1) Il est impossible de trouver une stratégie meilleure que η (a). Cela signifie qu'il n'existe aucune stratégie η' (a) telle que

$$L(\vartheta, \eta') \leqslant L(\vartheta, \eta) \qquad (9\text{-}6)$$

pour tous les $\vartheta \in \Theta$, bien que, pour certains ϑ, la relation (9-6) soit vraie. Dans ce cas, la stratégie η (a) peut être appelée *stratégie admissible*. Mais il n'est pas obligatoire que cette stratégie soit préférable, car d'autres stratégies peuvent aussi être prises en considération.

2) Il y a une stratégie η' (a) qui est meilleure que η (a). Cela veut dire que, pour la stratégie η' (a), la relation (9-6) sera vraie pour tous les $\vartheta \in \Theta$. Dans ce cas, la stratégie η (a) doit céder la place à la stratégie η' (a) et être considérée comme *stratégie inadmissible*.

Les stratégies admissibles sont commodes à examiner en termes d'un S-jeu. Etant donné que dans un S-jeu la stratégie du statisticien est définie par le point S de l'enveloppe convexe S^*, tandis que les pertes pour différents $\vartheta \in \Theta$ sont définies par les coordonnées de ce point, la stratégie définie par le point S sera une stratégie admissible

s'il n'y a aucun point $S' \in S^*$ dont toutes les coordonnées seront inférieures aux coordonnées respectives du point S.

Examinons la méthode qui permet de trouver les stratégies admissibles pour le cas où l'espace des états de la nature ne comporte que deux éléments ϑ_1 et ϑ_2. Sur la figure 9-2 est montré le domaine convexe S^* correspondant au cas considéré.

Examinons la stratégie définie par le point S_1 qui est un point intérieur du domaine S^*. Cette stratégie n'est pas une stratégie admissible, car tous les points situés sur le segment OS_1 à l'intérieur du domaine S^* définissent des stratégies meilleures que S_1. La meilleure de ces stratégies est la stratégie S qui appartient à la borne inférieure gauche du domaine S^*. Pour cette raison, tous les points intérieurs peuvent être éliminés en faveur des points constituant la frontière inférieure gauche du domaine S^* tracée sur la figure 9-2 en trait fort.

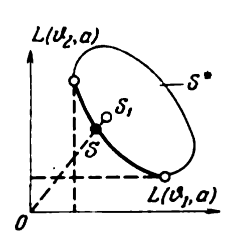


Fig. 9-2. Stratégies admissibles dans un S-jeu

Mais le déplacement du point le long de cette frontière n'apporte aucun avantage, car la diminution des pertes correspondant à un état de la nature est accompagnée de l'augmentation des pertes correspondant à l'autre état de la nature. C'est pourquoi les points appartenant à la frontière inférieure gauche du domaine S^* définissent justement les stratégies admissibles du statisticien.

Exemple 9-3. Dans le problème de la ligne technologique (voir fig. 9-1), la frontière inférieure gauche du domaine S^* comporte les segments C_1C_2 et C_2C_3 dont chacun est déterminé par un mélange de stratégies pures a_1, a_2 et a_2, a_3. Soit $0 \leqslant w \leqslant 1$. Alors l'équation du segment C_1C_2 s'écrira sous forme vectorielle comme suit:

$$S = wC_1 + (1 - w)\,C_2. \tag{9-7}$$

Cette équation définit la stratégie mixte $\eta(a) = (w, 1 - w, 0)$. En projetant l'équation du segment C_1C_2 sur les axes de coordonnées, on obtient:

$$\left.\begin{aligned} L(\vartheta_1, \eta) &= 0\cdot w + 1\,(1 - w) = 1 - w;\\ L(\vartheta_2, \eta) &= 5w + 3\,(1 - w) = 3 + 2w.\end{aligned}\right\} \tag{9-8}$$

D'une façon analogue, l'équation du segment C_2C_3 qui définit la stratégie mixte $\eta(a) = (0, w, 1 - w)$ se réduit à la forme

$$\left.\begin{aligned} L(\vartheta_1, \eta) &= 3 - 2w;\\ L(\vartheta_2, \eta) &= 2 + w.\end{aligned}\right\} \tag{9-9}$$

c) Principes de choix des stratégies dans les jeux statistiques

On appelle principe de choix une règle qui permet de déterminer la meilleure stratégie mixte du statisticien. En différentes circonstances, le statisticien peut faire appel aux différents principes de choix de sa stratégie.

Un des principes possibles du choix de la stratégie peut être le *principe du minimax*. Ce principe est utilisé avec profit dans les jeux stratégiques lorsqu'on a affaire à un adversaire raisonnable qui désire nous infliger la perte maximale. Quand même, dans certaines circonstances, il est également rationnel d'utiliser ce principe dans les jeux statistiques.

Conformément au principe du minimax, le statisticien doit choisir une stratégie mixte $\eta(a)$ telle que les pertes moyennes $L(\vartheta, \eta)$ soient minimales, l'état de la nature ϑ lui étant le plus défavorable. Le pire des cas se présentera lorsque $\vartheta \in \Theta$ va déterminer le maximum de la quantité $L(\vartheta, \eta)$. C'est justement cette quantité que le statisticien doit minimiser, c.-à-d. choisir la stratégie $\eta^*(a)$ qui assure la condition

$$L(\vartheta, \eta^*) \equiv \min_{\eta} \max_{\vartheta} L(\vartheta, \eta). \qquad (9\text{-}10)$$

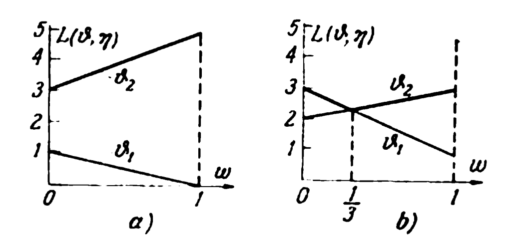


Fig. 9-3. Détermination de la stratégie optimale suivant le principe du minimax (a — correspond au segment C_1C_2; b — au segment C_2C_3 sur la figure 9-1)

Exemple 9-4. Trouvons la stratégie minimax dans le problème de la ligne technologique. Etant donné que la solution doit se trouver dans la classe des stratégies admissibles, on peut se borner à l'examen des stratégies définies par les relations (9-8) et (9-9) et correspondant aux segments C_1C_2 et C_2C_3 de l'enveloppe convexe S^* représentée sur la figure 9-1. En conformité avec ces relations et ces segments, sur les figures 9-3, a et 9-3, b on a construit les graphiques de variation $L(\vartheta, \eta)$ en fonction de w pour $\vartheta = \vartheta_1$ et $\vartheta = \vartheta_2$. Les valeurs de $\max_{\vartheta} L(\vartheta, \eta)$ sont tracées en trait fort. Les graphiques montrent que le minimum de cette grandeur sur le segment C_1C_2 est atteint pour $w = 0$ et est égal à 3; sur le segment C_2C_3 le minimum est défini par la condition d'intersection de deux droites

$$3 - 2w = 2 + w,$$

ce qui veut dire qu'il a lieu pour $w = 1/3$ et est égal à $7/3 < 3$. Ainsi, le principe du minimax nous donne le point situé sur le segment C_2C_3 et correspondant à $w = 1/3$, c.-à-d. définit la stratégie mixte optimale $\eta^* = (0, 1/3, 2/3)$ pour laquelle les pertes du statisticien ne dépasseront pas 7/3, quelle que soit la stratégie de la nature.

Parfois, il est rationnel de choisir la stratégie en partant des pertes dites ***supplémentaires*** $L'(\vartheta, a)$ au lieu de s'appuyer sur les pertes totales $L(\vartheta, a)$. Les pertes supplémentaires sont définies par la relation

$$L'(\vartheta, a) = L(\vartheta, a) - \min_{a} L(\vartheta, a). \qquad (9\text{-}11)$$

Pour chaque état de la nature, la grandeur $\min_a L(\vartheta, a)$ détermine les pertes minimales que le statisticien doit obligatoirement subir (pertes nécessaires), même si ses actions sont les meilleures. On peut considérer que les pertes nécessaires sont compensées d'une façon ou d'une autre (en fixant, par exemple, des prix adéquats pour les biens fabriqués) et, par conséquent, peuvent être négligées lors du choix de la stratégie. Dans ce cas, le choix de la stratégie peut se faire suivant le *principe du minimax des pertes supplémentaires.*

Exemple 9-5. Les pertes supplémentaires dans le problème de la ligne technologique trouvées à l'aide de la relation (9-11) sont résumées dans le tableau 9-3. En appliquant à ce tableau le principe du minimax, on obtient comme stratégie optimale la stratégie pure a_2 dont les pertes sont égales à 1.

Tableau 9-3

Pertes supplémentaires dans le problème de la ligne technologique

ϑ	A		
	a_1	a_2	a_3
ϑ_1	0	1	3
ϑ_2	3	1	0

Les principes du minimax, partant de l'hypothèse que la nature agit de la façon la plus défavorable pour le statisticien, se trouvent justifiés dans les jeux stratégiques, tandis que dans les jeux statistiques ils expriment, dans le fond, le point de vue d'une personne très prudente qui s'efforce d'obtenir au moins ce qui est accessible, sans tâcher d'atteindre le maximum pour ne pas subir accidentellement des pertes plus importantes. Un autre inconvénient des principes du minimax consiste dans le fait qu'ils ne tiennent pas compte de l'information a priori sur les états de la nature en limitant ainsi le gain que cette information pourrait fournir.

Pour cette raison, les principes du minimax peuvent être recommandés dans les cas où il n'y a pas d'information a priori sur les états de la nature ou bien si cette information est douteuse.

Un autre principe de choix de la stratégie qui prend en considération la distribution a priori de probabilités $\xi(\vartheta)$ est le *principe de Bayes.* Conformément à ce principe, l'estimation de la stratégie mixte $\eta(a)$ du statisticien se fait en prenant la moyenne des pertes $L(\vartheta, \eta)$ suivant tous les états possibles de la nature $\vartheta \in \Theta$ compte tenu de la distribution a priori de probabilités $\xi(\vartheta)$, c.-à-d. suivant la grandeur

$$L(\xi, \eta) = \sum_{\vartheta} L(\vartheta, \eta)\, \xi(\vartheta). \tag{9-12}$$

Dans ces conditions, la meilleure stratégie $\eta(a)$ sera celle qui assure le minimum de la grandeur $L(\xi, \eta)$. Elle sera appelée *stratégie de Bayes.*

Le principe de Bayes peut être, bien entendu, appliqué aussi bien aux pertes totales qu'aux pertes supplémentaires, mais, dans les exemples qui suivent, nous n'allons l'appliquer qu'aux pertes totales.

Exemple 9-6. Dans le problème de la ligne technologique, pour $\xi(\vartheta_1) = 0,6$ et $\xi(\vartheta_2) = 0,4$ utilisées pour les stratégies admissibles définies par le segment C_1C_2, on a:

$$L(\xi, \eta) = (1 - w)\,0,6 + (3 + 2w)\,0,4 = 1,8 + 2,0w,$$

de sorte que $\min_w L(\xi, \eta) = 1,8$ pour $w = 0$, ce qui correspond à la stratégie mixte $\eta = (0, 1, 0)$; pour le segment C_2C_3:

$$L(\xi, \eta) = 2,6 - 0,8w,$$

de sorte que $\min_w L(\xi, \eta) = 1,8$ pour $w = 1$, ce qui correspond à la même stratégie mixte $\eta = (0, 1, 0)$. De cette façon, la stratégie de Bayes est représentée par la stratégie pure a_2.

Il est à remarquer que le résultat obtenu n'est pas fortuit. Plus loin nous allons montrer que le principe de Bayes donne toujours comme la meilleure stratégie une des stratégies pures du statisticien, l'action de Bayes a^* étant définie par la condition (9-3).

d) Interprétation géométrique des stratégies de Bayes

Examinons un S-jeu statistique pour deux états de la nature ϑ_1 et ϑ_2 défini par le domaine convexe S^* du plan (x, y) représenté sur la figure 9-4, où

$$x = L(\vartheta_1, \eta); \quad y = L(\vartheta_2, \eta). \tag{9-13}$$

On sait que les stratégies admissibles se situent sur la frontière inférieure gauche du domaine S^*. Considérons une des stratégies admissibles S_0. Construisons l'ensemble auxiliaire R contenant tous les points situés au-dessous et à gauche du point S_0. Il est évident que S^* et R sont des ensembles convexes et qu'aucun point de S^* n'appartient à R, y compris le point S_0.

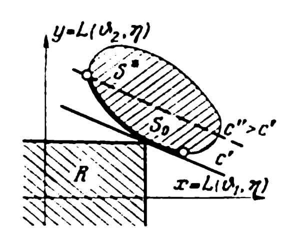


Fig. 9-4. Construction de la ligne d'appui à un ensemble convexe

Construisons la droite qui sépare les ensembles S^* et R. Cette droite doit passer par le point S_0, c.-à-d. doit être une droite d'appui à l'ensemble S^* au point S_0. Mais elle doit en même temps être une droite d'appui à l'ensemble R. Donc, elle doit avoir une pente négative, ou être verticale, ou bien être horizontale et son équation peut s'écrire

$$y = -kx + c, \quad k \geqslant 0 \tag{9-14}$$

En divisant les deux membres de cette équation par $k + 1$, ramenons-la à la forme

$$ax + by = c', \tag{9-15}$$

où

$$a = \frac{k}{k+1}; \quad b = \frac{1}{k+1}; \quad c' = \frac{c}{k+1}. \tag{9-16}$$

En observant que $a \geqslant 0$, $b \geqslant 0$, $a + b = 1$ on peut poser $a = w$, $b = 1 - w$ et considérer w et $1 - w$ en tant que distributions a priori des états de la nature ϑ_1 et ϑ_2:

$$w = \xi(\vartheta_1), \quad 1 - w = \xi(\vartheta_2). \tag{9-17}$$

Dans ce cas, l'équation de la droite d'appui peut s'écrire

$$wx + (1 - w)\, y = c' \tag{9-18}$$

ou

$$L(\vartheta_1, \eta)\, \xi(\vartheta_1) + L(\vartheta_2, \eta)\, \xi(\vartheta_2) = c'. \tag{9-19}$$

On voit donc que la grandeur c' détermine les pertes moyennes $L(\xi, \eta)$ pour les probabilités a priori d'états de la nature $\xi(\vartheta_1) = w$ et $\xi(\vartheta_2) = 1 - w$. D'autre part, il n'est pas difficile de remarquer que, pour le point S_0, la grandeur c' est minimale de toutes les valeurs possibles de $L(\xi, \eta)$, car si c' augmente jusqu'à $c'' > c'$, cela signifie que la droite va passer plus haut par des points appartenant à l'ensemble S^*, donc inadmissibles, tandis qu'une diminution de c' n'est pas possible vu que, dans ce cas, la droite passera plus bas et n'aura plus de points communs avec S^*. De cette façon, la grandeur c' définit la stratégie η qui donne le minimum des pertes moyennes $L(\xi, \eta)$ pour la distribution considérée $\xi(\vartheta) = (w, 1 - w)$, c.-à-d. définit la stratégie de Bayes pour la distribution a priori de probabilités donnée.

Etant donné que S_0 est un point arbitraire situé sur la frontière des stratégies admissibles, on peut trouver pour tout point de cette frontière des w et $1 - w$ tels que ce point définisse la stratégie de Bayes. Il s'ensuit que, *pour certaines probabilités a priori w et $1 - w$, chaque stratégie admissible est une stratégie de Bayes.*

Supposons maintenant données les probabilités a priori d'états de la nature w et $1 - w$; on demande de trouver, pour ce cas, le point de l'enveloppe convexe S^*, qui définit la stratégie de Bayes. Construisons sur le plan (x, y) la droite

$$wx + (1 - w)\, y = c. \tag{9-20}$$

Pour un c arbitraire, cette droite sera parallèle à la droite d'appui passant par le point correspondant à la stratégie de Bayes. Pour rendre la construction plus aisée, posons $c = w(1 - w)$ et écrivons l'équation (9-20) comme suit:

$$\frac{x}{1-w} + \frac{y}{w} = 1. \tag{9-21}$$

On obtient ainsi l'équation aux segments de la droite représentée sur la figure 9-5. Il est facile maintenant de construire la droite d'appui correspondant aux valeurs données de w et $1 - w$, droite qui définira sur l'enveloppe convexe S^* le point correspondant à la stratégie de Bayes du statisticien.

Vu que la frontière extérieure de l'enveloppe convexe représente un polygone dont les sommets correspondent aux stratégies pures du statisticien, la droite d'appui doit obligatoirement passer par au moins un de ces sommets. Donc, *pour les probabilités a priori données w et $1 - w$, il existe toujours au moins une stratégie de Bayes qui est une stratégie pure.* Cette circonstance simplifie beaucoup la résolution des jeux statistiques, car, en cherchant la solution de Bayes, on peut limiter la recherche à un nombre fini de stratégies admissibles pures au lieu d'examiner une infinité de stratégies mixtes. Cela arrange aussi beaucoup de statisticiens sceptiques qui hésitent d'utiliser les stratégies mixtes en prétextant qu'elles impliquent la mise en jeu d'un mécanisme aléatoire qui n'a aucun rapport au fond du problème.

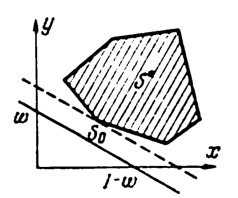


Fig. 9-5. Détermination géométrique de la stratégie de Bayes

Bien entendu, toutes les conclusions tirées restent vraies pour le cas où il y a plus de deux états de la nature, mais l'interprétation géométrique devient alors plus difficile, car l'examen du problème dans le plan est impossible.

9-3. JEUX STATISTIQUES À EXPÉRIENCE UNIQUE

a) Position du problème

Comme on l'a déjà remarqué auparavant, une des particularités des jeux statistiques consiste dans le fait que le statisticien peut élargir et préciser ses connaissances relatives à l'état de la nature en faisant des expériences. En principe, si les expériences pouvaient être continuées sans aucune limitation, le statisticien pourrait obtenir une information complète sur l'état de la nature, ce qui lui permettrait d'agir en toute netteté. Pourtant, les frais entraînés par les expériences peuvent s'avérer supérieurs au gain qu'elles apporteraient.

La possibilité de faire des expériences élargit énormément la classe de stratégies du statisticien. Avant tout, ce dernier doit décider si les expériences doivent être faites ou non. Ensuite, il doit déterminer la nature des expériences, leur nombre et les actions qu'il faut entreprendre en fonction des résultats obtenus.

Aux paragraphes suivants, nous tâcherons de répondre à certaines de ces questions, le présent paragraphe étant consacré aux jeux statistiques à expérience unique.

Pour le moment, nous n'allons pas inclure dans la classe des stratégies du statisticien la prise de décision sur les expériences en considérant que la décision de faire une expérience unique est déjà

prise. Avec cela, nous allons entendre par expérience unique une expérience dont le volume et l'ordre d'exécution sont définis à l'avance. Ainsi, s'il faut vérifier si une pièce de monnaie est symétrique, on peut faire une expérience unique comportant n jets de cette pièce. Pour $n = 5$, l'issue de l'expérience peut être exprimée par la suite *PFPPF*. En tout, il y a 2^n suites de ce type, ce qui signifie que, dans ce cas, l'espace des issues de l'expérience comporte 2^n éléments. D'une façon analogue, lorsqu'on vérifie une arme, on entend par expérience unique une expérience où l'on tire n coups. Pour estimer l'influence exercée par une certaine nourriture spéciale sur un animal, on peut faire une expérience unique qui consiste à mesurer journellement pendant plusieurs mois l'augmentation de poids chez n animaux, etc. Bien que l'expérience des exemples cités comprenne une série d'épreuves, nous l'appelons expérience unique étant donné que le nombre d'épreuves et le caractère de chacune d'elles sont définis à l'avance.

b) Espace d'échantillonnage

Désignons par Z l'espace des issues de l'expérience et par $z_1, \ldots$ $\ldots, z_\nu$ les éléments de cet espace (issues isolées de l'expérience).

Les issues isolées $z \in Z$ de l'expérience sont liées aux états de la nature $\vartheta \in \Theta$ par le fait qu'à chaque état de la nature correspond une probabilité déterminée $p_\vartheta(z)$ que l'issue de l'expérience sera l'issue $z \in Z$ donnée. Les grandeurs $p_\vartheta(z)$, notées encore $p(z \mid \vartheta)$, représentent la distribution conditionnelle de probabilités sur l'espace Z pour un ϑ donné et satisfont aux relations

$$p_\vartheta(z) \geqslant 0; \quad \sum_z p_\vartheta(z) = 1. \tag{9-22}$$

La réunion de trois éléments — espace des issues de l'expérience Z, espace des états de la nature Θ et distribution de probabilités $p_\vartheta(z)$ sur l'espace Z pour un $\vartheta \in \Theta$ donné — est appelée *espace d'échantillonnage* et se note

$$\mathfrak{Z} = (Z, \Theta, p) \tag{9-23}$$

Il est commode de donner l'espace d'échantillonnage sous forme d'un tableau contenant la distribution $p_\vartheta(z)$ sur le produit direct des ensembles $\Theta \times Z$.

Exemple 9-7. Dans le problème du remplacement de l'équipement, l'expérience peut consister en essais de contrôle de l'équipement par les moyens de l'entreprise. Dans ce cas, le manque d'un personnel hautement qualifié et d'appareils de contrôle et de mesure nécessaires fait que les résultats des essais ne reflètent qu'approximativement l'état réel de l'équipement. Supposons que l'expérience puisse avoir quatre issues possibles:

z_1: l'équipement est en bon état;

z_2: la réparation courante est nécessaire;

z_3: il faut remplacer les pièces usées;
z_4: l'équipement ne peut plus rester en service.
Les probabilités de chacune de ces issues pour différents états de la nature sont données au tableau 9-4 qui représente l'espace d'échantillonnage du problème en question.

Tableau 9-4

Espace d'échantillonnage dans le problème du remplacement de l'équipement

ϑ	z			
	z_1	z_2	z_3	z_4
ϑ_1	1/2	1/2	0	0
ϑ_2	0	1/2	1/2	0
ϑ_3	0	0	1/3	2/3

Tableau 9-5

Espace d'échantillonnage dans le problème de la ligne technologique

ϑ	z		
	z_1	z_2	z_3
ϑ_1	0,60	0,25	0,15
ϑ_2	0,20	0,30	0,50

Exemple 9-8. Dans le problème de la ligne technologique, l'expérience peut prendre la forme d'une analyse préalable grossière de la teneur en impuretés (l'analyse de précision effectuée en laboratoire est exclue, car elle nécessite beaucoup de temps pendant lequel l'équipement doit rester en inactivité). Les résultats de l'expérience sont les suivants:
z_1: il n'y a pas d'impuretés;
z_2: il y a peu d'impuretés;
z_3: il y a beaucoup d'impuretés.
Dans ce cas, l'espace d'échantillonnage peut se présenter comme indiqué au tableau 9-5.

c) Fonction de décision

Dans un problème sans expérience, le statisticien doit prendre une décision appartenant à l'espace des décisions A en se basant sur l'information a priori $\xi(\vartheta)$ concernant les états de la nature. Dans un problème à expérience, la décision prise par le statisticien est fonction de l'issue $z \in Z$ de l'expérience. Pour formaliser ce problème, il peut analyser à l'avance toutes les issues possibles de l'expérience et élaborer une règle d qui détermine quelle décision $a \in A$ il faut prendre pour chacune des issues possibles $z \in Z$ de l'expérience. Cette règle représentera une application de l'espace des issues Z de l'expérience sur l'espace des décisions A

$$d: Z \to A, \tag{9-24}$$

ce qui peut encore s'écrire

$$d(z) = a.$$

La règle $d(z)$, qui détermine la décision $a \in A$ que le statisticien doit prendre quelle que soit l'issue $z \in Z$ de l'expérience, est appelée *fonction de décision.*

Expliquons la notion de fonction de décision sur l'exemple suivant. Supposons que l'espace des décisions comprenne trois éléments $A = \{a_1, a_2, a_3\}$ et que l'espace des issues de l'expérience inclut cinq éléments $Z = \{z_1, z_2, z_3, z_4, z_5\}$. La fonction de décision $d(z_i) = a_i$ peut être donnée sous la forme d'un ensemble des couples d'indices (i, j) définissant la décision a_j pour l'issue z_i de l'expérience. Une fonction de décision sera, par exemple, représentée par l'ensemble $\{(1, 1), (2,1), (3, 2), (4, 2), (5, 3)\}$ qui signifie que, pour les issues z_1 et z_2, on prend la décision a_1, pour les issues z_3 et z_4, la décision a_2, et pour l'issue z_5, la décision a_3. Certes, la fonction de décision citée n'est pas l'unique possible. On pourrait aussi bien considérer les fonctions de décision de la forme $\{(1, 1,), (2, 2), (3, 2), (4, 3), (5, 3)\}$ ou $\{(1, 3), (2, 1), (3, 2), (4, 3), (5, 2)\}$, etc. Compte tenu de ce qui vient d'être dit, il est commode d'introduire l'espace D comportant la totalité des fonctions de décision possibles et de l'appeler *espace des fonctions de décision.*

La fonction de décision considérée décompose l'ensemble Z en sous-ensembles disjoints S_{a_i}, notamment

$$S_{a_1} = \{z_1, z_2\}, \quad S_{a_2} = \{z_3, z_4\}, \quad S_{a_3} = \{z_5\}. \tag{9-25}$$

En utilisant la terminologie de la théorie des ensembles, on peut dire que toute fonction de décision $d \in D$ peut être considérée comme la décomposition de l'espace des issues de l'expérience Z en sous-ensembles disjoints S_a tels que

$$S_a = \{z: d(z) = a\}, \quad a \in A. \tag{9-26}$$

Si, en particulier, l'espace des décisions A ne comporte que deux éléments a_1 et a_2 (problème bialternatif), la fonction de décision $d(z)$ décompose l'espace Z en l'espace S (appelé domaine critique) et en son complémentaire $C(S) = \bar{S}$ définis par les conditions

$$d(z) = \begin{cases} a_1 & \text{si } z \in S; \\ a_2 & \text{si } z \in C(S). \end{cases} \tag{9-27}$$

La notion de fonction de décision permet de formuler d'une façon plus nette le problème du statisticien. Ce problème consiste à choisir dans l'espace des fonctions de décision D une fonction de décision $d(z)$ telle qu'il soit possible de prendre la décision la plus avantageuse. Pourtant, pour ce faire, il faut savoir estimer les différentes fonctions de décision. On y parvient à l'aide de la fonction de risque.

d) Fonction de risque

Si le statisticien a arrêté son choix sur une certaine fonction de décision $d(z)$, il a déterminé par là même pour chaque issue de l'expérience $z \in Z$ la décision $a = d(z)$ à laquelle, pour $\vartheta \in \Theta$ donné,

correspondront les pertes

$$L(\vartheta, a) = L[\vartheta, d(z)] = L_z(\vartheta, d). \tag{9-28}$$

Toutefois, pour ϑ donné, l'issue z de l'expérience sera une variable aléatoire déterminée par la distribution de probabilités $p_\vartheta(z)$ sur l'espace Z. Il s'ensuit donc que, pour z donné, les pertes $L_z(\vartheta, d)$ auront lieu avec la même probabilité $p_\vartheta(z)$ et représenteront aussi une variable aléatoire.

L'estimation de la fonction de décision $d(z)$ pour un état de la nature ϑ donné exigeant de prendre en considération toutes les issues possibles de l'expérience, il est nécessaire de considérer les pertes moyennes déterminees sur tout l'espace des issues de l'expérience Z. Ces pertes moyennes sont appelées *fonction de risque*. Elles sont notées $\rho(\vartheta, d)$ et sont déterminées à partir de la relation

$$\rho(\vartheta, d) = M_z[L_z(\vartheta, d)] = \sum_z L_z(\vartheta, d)\, p_\vartheta(z). \tag{9-29}$$

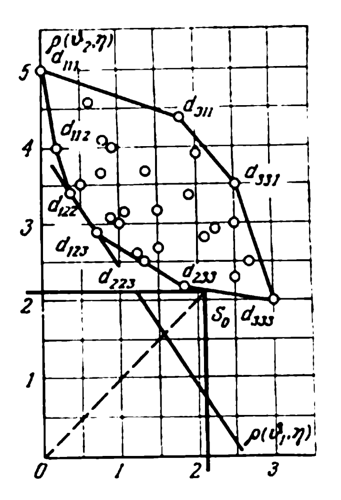


Fig. 9-6. Fonction de risque dans le problème de la ligne technologique

A chaque état de la nature $\vartheta \in \Theta$ et à chaque fonction de décision $d \in D$ sera associée une valeur bien déterminée des pertes moyennes, c.-à-d. de la fonction de risque $\rho(\vartheta, d)$ qui, de cette façon, est déterminée sur le produit direct des ensembles $\Theta \times D$ d'une manière absolument analogue à la détermination de la fonction de pertes $L(\vartheta, a)$ sur le produit direct des ensembles $\Theta \times A$ dans un jeu sans expérience. On arrive ainsi à la conclusion que l'espace des fonctions de décision D et la fonction de risque $\rho(\vartheta, d)$ jouent le même rôle dans un jeu à expérience unique que l'espace des décisions A et la fonction de pertes dans un jeu sans expérience. D'ici, des méthodes analogues de résolution de ces deux problèmes.

Dans un jeu à expérience unique, le statisticien peut aussi adopter des stratégies mixtes. A cette fin, il doit posséder un mécanisme aléatoire pour définir la distribution de probabilités $\eta(d)$ sur l'espace D. Lors de l'utilisation de la stratégie mixte $\eta(d)$, la fonction de risque, désignée dans ce cas par $\rho(\vartheta, \eta)$, s'obtient en prenant la moyenne de $\rho(\vartheta, d)$ suivant toutes les stratégies pures qui font partie de la stratégie mixte donnée. Ainsi donc,

$$\rho(\vartheta, \eta) = M_d[\rho(\vartheta, d)] = \sum_d \rho(\vartheta, d)\, \eta(d) \tag{9-30}$$

ou bien, compte tenu de (9-29),

$$\rho(\vartheta, \eta) = \sum_{z, d} L_z(\vartheta, d)\, p_\vartheta(z)\, \eta(d). \qquad (9\text{-}31)$$

Il est bien naturel qu'en cherchant sa meilleure stratégie dans un jeu à expérience unique, le statisticien ne doit partir que des stratégies admissibles qui sont déterminées d'une façon absolument analogue au cas d'un jeu sans expérience.

Exemple 9-9. Déterminons la fonction de risque dans le problème de la ligne technologique. Pour rendre les calculs plus commodes, portons dans un même tableau 9-6 les pertes du statisticien $L(\vartheta, a)$ et les probabilités des issues de l'expérience $p_\vartheta(z)$ données respectivement par les tableaux 9-2 et 9-5.

Tableau 9-6

Pertes et probabilités des issues de l'expérience dans le problème de la ligne technologique

ϑ	a			$p_\vartheta(z)$		
	a_1	a_2	a_3	z_1	z_2	z_3
ϑ_1	0	1	3	0,60	0,25	0,15
ϑ_2	5	3	2	0,20	0,30	0,50

Tableau 9-7

Valeurs de la fonction de risque $\rho(\vartheta, d)$ dans le problème de la ligne technologique

ϑ	d_{111}	d_{112}	d_{113}	d_{121}	d_{122}	d_{123}	d_{131}	d_{132}	d_{133}
ϑ_1	0,00	0,15	0,45	0,25	0,40	0,70	0,75	0,90	1,20
ϑ_2	5,00	4,00	3,50	4,40	3,40	2,90	4,10	3,10	2,60

ϑ	d_{211}	d_{212}	d_{213}	d_{221}	d_{222}	d_{223}	d_{231}	d_{232}	d_{233}
ϑ_1	0,60	0,75	1,05	0,85	1,00	1,30	1,35	1,50	1,80
ϑ_2	4,60	3,60	3,10	4,00	3,00	2,50	3,70	2,70	2,20

ϑ	d_{311}	d_{312}	d_{313}	d_{321}	d_{322}	d_{323}	d_{331}	d_{332}	d_{333}
ϑ_1	1,80	1,95	2,25	2,05	2,20	2,50	2,55	2,70	3,00
ϑ_2	4,40	3,40	2,90	3,80	2,80	2,30	3,50	2,50	2,00

L'espace des issues de l'expérience comportant trois éléments, la fonction de décision sera de la forme $d(z) = (a_i, a_j, a_k) = d_{ijk}$, où a_i, a_j et a_k sont les décisions que le statisticien doit prendre respectivement pour les issues de l'expérience z_1, z_2 et z_3. Ainsi, la fonction de décision d_{122} signifie que pour les issues z_1, z_2 et z_3 de l'expérience, le statisticien prend les décisions a_1, a_2, a_3.

Les valeurs de la fonction de risque calculées d'après les formules (9-29) pour chaque fonction de décision sont données au tableau 9-7 et sur la figure 9-6. Examinons la façon dont on a obtenu les données de ce tableau en refaisant les calculs relatifs à $\rho(\vartheta, d_{122})$. Conformément à (9-29), on a:

$$\rho(\vartheta, d_{122}) = L_{z_1}(\vartheta, d_{122})\, p_\vartheta(z_1) + L_{z_2}(\vartheta, d_{122})\, p_\vartheta(z_2) + L_{z_3}(\vartheta, d_{122})\, p_\vartheta(z_3) = L(\vartheta, a_1)\, p_\vartheta(z_1) + L(\vartheta, a_2)\, p_\vartheta(z_2) + L(\vartheta, a_2)\, p_\vartheta(z_3).$$

En supposant $\vartheta = \vartheta_1$ et $\vartheta = \vartheta_2$, on obtient:

$$\rho(\vartheta_1, d_{122}) = 0{,}4; \quad \rho(\vartheta_2, d_{122}) = 3{,}4.$$

e) Principes de choix de la stratégie dans les jeux à expérience unique

Vu que l'introduction de la fonction de risque ramène un jeu à expérience unique à une forme analogue à un jeu sans expérience tous les principes de choix de la stratégie dans un jeu sans expérience restent valables à une seule différence près qu'au lieu de minimiser les pertes moyennes, le statisticien doit maintenant minimiser le risque moyen.

Le principe du minimax consiste à choisir la stratégie $\eta(d)$ pour laquelle le risque moyen $\rho(\vartheta, \eta)$ soit minimal dans le cas où l'état de la nature est le plus défavorable pour le statisticien. Cela signifie que la stratégie minimax η^* est choisie en partant de la condition

$$\rho(\vartheta, \eta^*) = \min_\eta \max_\vartheta \rho(\vartheta, \eta). \qquad (9\text{-}32)$$

Si en déterminant le risque moyen $\rho(\vartheta, \eta)$ on prend en considération les pertes supplémentaires $L'(\vartheta, a)$ au lieu des pertes totales, la formule (9-32) déterminera le principe du minimax des pertes supplémentaires.

Pour utiliser le principe de Bayes, introduisons la notion de *risque espéré* qui signifie le risque moyen compte tenu de tous les états possibles de la nature $\vartheta \in \Theta$ et de la distribution a priori de probabilités sur l'espace Θ. Etant donné que, lors de l'utilisation du principe de Bayes, le statisticien peut se limiter à l'adoption exclusive des stratégies pures, le risque espéré sera

$$\rho(\xi, d) = \sum_\vartheta \rho(\vartheta, d)\, \xi(\vartheta). \qquad (9\text{-}33)$$

Le principe de Bayes exige l'utilisation d'une fonction de décision d^* telle que le risque espéré, appelé dans ce cas *risque de Bayes*, soit minimal:

$$\rho^*(\xi) = \rho(\xi, d^*) = \min_d \rho(\xi, d). \qquad (9\text{-}34)$$

Exemple 9-10. Déterminons la stratégie minimax et la stratégie de Bayes dans le problème de la ligne technologique à expérience unique. Sur la figure 9-6 ce problème est représenté sous la forme d'un S-jeu. La construction géométrique permet de voir sans difficulté que la stratégie minimax est définie par le point S_0 de coordonnées (30/14, 30/14) et correspond à l'adoption des stratégies pures d_{233} et d_{333} avec les probabilités 10/14 et 4/14. La stratégie de Bayes sera la stratégie d_{123}.

9-4. UTILISATION DES PROBABILITÉS A POSTERIORI

a) Détermination du nombre de stratégies dans les jeux avec expérience

Les exemples des paragraphes précédents montrent que les expériences faites dans les jeux statistiques entraînent une augmentation considérable du nombre de stratégies pures du statisticien. Ainsi, dans le cas du problème de la ligne technologique, le nombre de stratégies pures augmente de 3 à 27 quand on tient compte des résultats de l'expérience. Si l'on n'était pas parvenu à représenter ce jeu sous la forme d'un S-jeu dans le plan, on se serait heurté à des difficultés considérables lors de son analyse.

Désignons par N le nombre total de stratégies pures du statisticien dans un jeu à expérience unique. Ce nombre est facile à calculer. Supposons que dans un jeu sans expérience le statisticien dispose de l décisions possibles $a_1, \ldots, a_l$, tandis que le nombre d'issues possibles de l'expérience est ν.

Pour $\nu = 1$, on a un jeu dont l'issue de l'expérience est connue à l'avance, ce qui équivaut à un jeu sans expérience à $N = l$.

Pour $\nu = 2$, la stratégie du statisticien peut s'écrire sous la forme $d = (a_{i_1}, a_{i_2})$, où les quantités a_{i_1} et a_{i_2} peuvent être représentées par n'importe quels $a \in A$. On peut donc avoir l valeurs différentes de a_{i_1} et à chacune de ces valeurs on peut mettre en correspondance l valeurs différentes de a_{i_2} de sorte que $N = l^2$.

Pour $\nu = 3$, la stratégie du statisticien est $d = (a_{i_1}, a_{i_2}, a_{i_3})$. L'ensemble des solutions a_{i_1} et a_{i_2} peut prendre l^2 valeurs à chacune desquelles on peut faire correspondre l valeurs de a_{i_3}, de sorte que $N = l^3$.

Par un raisonnement analogue on arrive à la conclusion que pour l et ν arbitraires, le nombre total de stratégies pures du statisticien $N = l^\nu$. Pour des l et ν grands, le nombre de stratégies pures du statisticien peut devenir si important que leur analyse rencontre des difficultés majeures.

Dans ces cas, la recherche des stratégies de Bayes peut être simplifiée de beaucoup si l'on utilise au lieu de la distribution a priori de probabilités la distribution a posteriori calculée sur la base des résultats de l'expérience effectuée. Le nombre de stratégies pures du statisticien dans le problème avec expérience reste alors le même que dans le problème sans expérience.

b) Distribution a posteriori de probabilités. Formule de Bayes

La distribution a priori de probabilités $\xi(\vartheta)$ obtenue sur la base des données statistiques et de l'expérience passée nous fournit une information utile sur la fréquence d'apparition de tel ou tel état de la nature, sans aucun rapport aux conditions concrètes dans lesquelles le statisticien doit prendre sa décision. Les expériences faites par le statisticien ont pour but d'apporter une information supplémentaire sur l'état réel de la nature.

Supposons que l'espace des issues de l'expérience soit donné par l'ensemble $Z = \{z_1, \ldots, z_\nu\}$. Avec cela, l'issue d'une expérience concrète est aléatoire et ne donne pas la possibilité de tirer des conclusions précises sur l'état réel de la nature. Toutefois, l'indétermination concernant l'état de la nature diminue sensiblement si l'expérience est correctement faite.

Cette variation de l'indétermination concernant l'état de la nature consiste dans le fait qu'à la suite de l'expérience on obtient à la place de la distribution a priori $\xi(\vartheta)$ une nouvelle distribution de probabilités $\xi_z(\vartheta)$ qui est appelée *distribution a posteriori de probabilités* sur l'espace Θ pour une issue concrète donnée $z \in Z$ de l'expérience. Voyons maintenant les méthodes de calcul de la distribution a posteriori de probabilités.

Remarquons tout d'abord que l'issue de l'expérience, dont le but consiste à préciser l'état réel de la nature, sera fonction de l'état de la nature ϑ. L'étude préalable des conditions dans lesquelles est feite l'expérience permet d'indiquer pour chaque état de la nature ϑ la distribution de probabilités sur l'espace des issues Z de l'expérience, qui, de ce fait, représente la distribution conditionnelle de probabilités $p(z \mid \vartheta) = p_\vartheta(z)$.

Pour décrire entièrement une expérience concrète, il faut savoir l'issue $z \in Z$ de l'expérience et l'état de la nature $\vartheta \in \Theta$ pour lequel l'expériense donnée a été faite. Ainsi donc, le résultat d'une expérience concrète peut être représenté comme un couple ordonné (z, ϑ) qui est un élément du produit direct des ensembles $Z \times \Theta$.

Désignons par $q(z, \vartheta)$ la distribution de probabilités sur l'ensemble $Z \times \Theta$. Dans le cas général, cette distribution de probabilités variera avec la variation de la distribution a priori de probabilités $\xi(\vartheta)$, c.-à-d. sera fonction de ξ. Pour une distribution a priori concrète $\xi(\vartheta)$, désignons la distribution de probabilités sur l'ensemble $Z \times \Theta$ par $q_\xi(z, \vartheta)$.

Il nous faut maintenant lier entre elles les distributions $\xi(\vartheta)$, $p_\vartheta(z)$, $\xi_z(\vartheta)$ et $q_\xi(z, \vartheta)$. Un problème de ce type s'est déjà présenté lors de l'examen des variables aléatoires à deux dimensions quand l'issue de l'expérience z était le couple ordonné (x, y).

Pour ces variables aléatoires, on a eu la relation (5-37) qui, appliquée à notre cas, peut s'écrire

$$q_{\xi}(z,\ \vartheta) = p_{\vartheta}(z)\ \xi(\vartheta) = \xi_z(\vartheta)\ p(z). \qquad (9\text{-}35)$$

D'ici on trouve:

$$\xi_z(\vartheta) = \frac{p_{\vartheta}(z)\ \xi(\vartheta)}{p(z)}. \qquad (9\text{-}36)$$

Dans cette expression, $p(z)$ représente la probabilité inconditionnelle de l'issue donnée z de l'expérience, c.-à-d. la probabilité pour que l'issue z ait lieu pour un état arbitraire de la nature. Compte tenu de ce fait, la probabilité $p(z)$ peut être mise sous la forme

$$p(z) = p(z \mid \vartheta_1 \cup \vartheta_2 \cup \ldots \cup \vartheta_m). \qquad (9\text{-}37)$$

En appliquant à cette expression la formule de la probabilité totale (5-42), on obtient:

$$p(z) = \sum_{\vartheta} p_{\vartheta}(z)\ \xi(\vartheta). \qquad (9\text{-}38)$$

En prenant en considération (9-38), la formule de la distribution a posteriori de probabilités $\xi_z(\vartheta)$ prend la forme

$$\xi_z(\vartheta) = \frac{p_{\vartheta}(z)\ \xi(\vartheta)}{\sum_{\vartheta} p_{\vartheta}(z)\ \xi(\vartheta)}. \qquad (9\text{-}39)$$

Dans la théorie des probabilités, cette formule est appelée *formule de Bayes.*

Exemple 9-11. Déterminons la distribution a posteriori de probabilités dans le problème de la ligne technologique. Les calculs effectués à l'aide de la formule (9-39) sont résumés au tableau 9-8.

Tableau 9-8

Calcul de la distribution a posteriori de probabilités dans le problème de la ligne technologique

ϑ	$\xi(\vartheta)$	$p_{\vartheta}(z)$			$p_{\vartheta}(z)\ \xi(\vartheta)$			$\xi_z(\vartheta)$		
		z_1	z_2	z_3	z_1	z_2	z_3	z_1	z_2	z_3
ϑ_1	0,6	0,60	0,25	0,15	0,36	0,15	0,09	0,818	0,555	0,310
ϑ_2	0,4	0,20	0,30	0,50	0,08	0,12	0,20	0,182	0,445	0,690
$\sum_{\vartheta} p_{\vartheta}(z)\ \xi(\vartheta)$					0,44	0,27	0,29			

Le calcul est effectué en partant de la distribution a priori de probabilités $\xi(\vartheta)$ donnée au tableau 9-2 et de la distribution conditionnelle de probabilités $p_\vartheta(z)$ prise du tableau 9-5.

Les valeurs des probabilités a posteriori trouvées montrent que, pour l'issue z_1 ou z_3 de l'expérience, l'indétermination relative aux états de la nature diminue considérablement. Mais pour l'issue z_2 de l'expérience, l'indétermination augmente par rapport à ce qu'elle était avant l'expérience.

c) Principe du maximum de vraisemblance

La connaissance des probabilités a posteriori permet d'estimer l'état de la nature en utilisant le principe du maximum de vraisemblance. Conformément à ce principe, on prend comme estimation de l'état de la nature $\hat{\vartheta}$ l'état de la nature qui semble être le plus probable en vertu des données expérimentales.

Exemple 9-12. Etablissons l'estimation de l'état de la nature $\hat{\vartheta}$ dans le problème de la ligne technologique pour l'issue z_1 de l'expérience. Conformément au tableau 9-8, les probabilités a posteriori des états de la nature ϑ_1 et ϑ_2 pour l'issue z_1 de l'expérience sont les suivantes:

$$\xi_{z_1}(\vartheta_1)=0{,}818; \quad \xi_{z_1}(\vartheta_2)=0{,}182.$$

Etant donné que $\max[\xi_{z_1}(\vartheta_1), \xi_{z_1}(\vartheta_2)] = \xi_{z_1}(\vartheta_1)$, d'après le principe de la vraisemblance maximale, on a $\hat{\vartheta} = \vartheta_1$.

Le principe du maximum de vraisemblance est souvent utilisé pour choisir la décision dans *un problème bialternatif.* Dans ce problème, l'espace des états de la nature comporte deux éléments $\Theta = \{\vartheta_1, \vartheta_2\}$ avec la distribution a priori de probabilités $\xi(\vartheta) = (\zeta, 1 - \zeta)$. L'espace des décisions du statisticien comprend également deux éléments $A = \{a_1, a_2\}$, où a_1 est la décision que l'état de la nature est ϑ_1; a_2 la décision que l'etat de la nature est ϑ_2. On connaît également les distributions conditionnelles de probabilités des issues de l'expérience $p_\vartheta(z)$ pour $\vartheta = \vartheta_1$ et pour $\vartheta=\vartheta_2$. Prendre la décision $a \in \{a_1, a_2\}$ d'après les résultats de l'issue $z \in Z$ de l'expérience.

Exemple 9-13. Problème du radar. Sur l'écran d'un radar un écho peut apparaître soit à la suite de la présence d'un objectif dans la zone de détection, soit par suite de l'action de différents parasites. Il s'ensuit donc que l'on peut se trouver en présence de deux états de la nature: ϑ_1 qui correspond à la présence de l'objectif, et ϑ_2 lorsque l'objectif est absent. En observant l'écran, on peut prendre deux décisions: a_1 qui suppose la présence de l'objectif, et a_2 qui affirme son absence. Lorsqu'on agit de cette façon, on peut commettre des erreurs de deux genres:

$a_1 \mid \vartheta_2$: erreur de premier genre appelée parfois « fausse alerte »;

$a_2 \mid \vartheta_1$: erreur de deuxième genre appelée parfois « manque d'objectif ».

Pour prendre la décision dans un problème bialternatif, on utilise souvent le *rapport de vraisemblance* défini par la relation

$$\Lambda(z)=\frac{\xi_z(\vartheta_1)}{\xi_z(\vartheta_2)}. \qquad (9\text{-}40)$$

On dit que l'on effectue la vérification suivant le rapport de vraisemblance si un nombre k est donné de façon que la décision soit prise en conformité de la règle suivante:

on prend la décision a_1 si $\Lambda(z) > k$;
on prend la décision a_2 si $\Lambda(z) < k$;
on prend la décision a_1 ou a_2 si $\Lambda(z) = k$.

La valeur de k est choisie en fonction des conséquences que peut avoir une décision erronée. Ainsi, dans le problème du poste radar, l'erreur du type « fausse alerte » peut avoir des conséquences beaucoup plus sérieuses que l'erreur du type « manque d'objectif ». La décision a_1 doit être prise seulement si l'on a la certitude que l'objectif existe réellement, c.-à-d. si la condition $\xi_z(\vartheta_1) \gg \xi_z(\vartheta_2)$ est satisfaite. Cela signifie que, dans ce cas, il faut prendre $k \gg 1$.

d) Détermination de la décision de Bayes par l'utilisation des probabilités a posteriori

Au début du présent chapitre, on a vu que la difficulté principale rencontrée lors de la résolution d'un problème avec expérience consistait dans l'augmentation brusque du nombre de stratégies avec l'augmentation du nombre d'issues possibles de l'expérience. Mais en réalité, il n'est pas du tout nécessaire de prendre en considération toutes les issues possibles de l'expérience. Si l'expérience aboutit à une certaine issue concrète $z \in Z$, le problème doit être résolu justement pour cette issue.

On peut le faire en calculant, pour l'issue donnée z de l'expérience, la distribution a posteriori de probabilités $\xi_z(\vartheta)$ sur l'espace des états de la nature Θ. Dans ce cas, on connaît l'espace des états de la nature Θ, l'espace des décisions A et la distribution de probabilités sur l'espace des états de la nature $\xi_z(\vartheta)$ compte tenu du résultat de l'expérience. Mais ce problème diffère du problème sans expérience seulement par l'utilisation de la distribution a posteriori de probabilités $\xi_z(\vartheta)$ au lieu de la distribution a priori $\xi(\vartheta)$. Donc, les méthodes de résolution de ce problème sont aussi analogues à celles adoptées dans le problème sans expérience.

Lors de l'utilisation des probabilités a posteriori $\xi_z(\vartheta)$ on prend en qualité de stratégies pures du statisticien les éléments de l'espace des décisions $A = \{a_1, \ldots, a_l\}$. Dans ce cas, chaque action $a \in A$ sera accompagnée des pertes du statisticien $L(\xi_z, a)$ déterminées, par analogie à (9-2), comme suit:

$$L(\xi_z, a) = M_{\vartheta|z}[Z(\vartheta, a)] = \sum_{\vartheta} L(\vartheta, a)\,\xi_z(\vartheta), \tag{9-41}$$

ou bien, compte tenu de (9-39),

$$L(\xi_z, a) = \frac{\sum_{\vartheta} L(\vartheta, a)\, p_{\vartheta}(z)\, \xi(\vartheta)}{\sum_{\vartheta} p_{\vartheta}(z)\, \xi(\vartheta)}. \tag{9-42}$$

Le principe de Bayes consiste à choisir une action $a^* \in A$ telle que les pertes définies d'après (9-42) soient minimales:

$$R^*(\xi_z) = R(\xi_z, a^*) = \min_a L(\xi_z, a). \tag{9-43}$$

e) Problème bialternatif

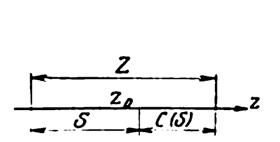
Illustrons les principes exposés en résolvant un problème bialternatif à distribution a priori de probabilités $\xi(\vartheta) = (\zeta, 1 - \zeta)$. Si les décisions sont correctes les pertes seront considérées comme nulles. L'erreur du premier genre $(a_2 \mid \vartheta_1)$ entraîne les pertes w, tandis que l'erreur du deuxième genre $(a_1 \mid \vartheta_2)$ donne comme pertes 1. Dans ce cas, la matrice de paiement prend la forme du tableau 9-9.

Tableau 9-9

Pertes dans un problème bialternatif

ϑ	a_1	a_2
ϑ_1	0	w
ϑ_2	1	0

Examinons la fonction de décision $d(z)$ qui, conformément à (9-27), décompose l'espace des issues Z de l'expérience en les domaines S et $C(S)$ tels que l'on prenne la décision a_1 si $z \in S$, et la décision a_2 si $z \in C(S)$. Etant donné que S et $C(S)$ doivent être compacts, la résolution du problème se réduit à la recherche de la frontière qui décompose l'ensemble Z en sous-ensembles disjoints S et $C(S)$. Désignons par z_0 les éléments $z \in Z$ qui constituent cette



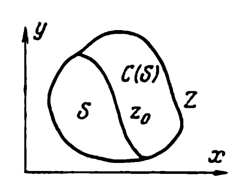


Fig. 9-7. Décomposition de l'espace des issues de l'expérience dans un problème bialternatif

frontière. Sur la figure 9-7 sont montrées les frontières z_0 pour les cas unidimensionnel et bidimensionnel.

Pour trouver l'équation qui définit la frontière z_0, écrivons les expressions des pertes moyennes pour z donné et pour les décisions a_1 et a_2. Compte tenu de (9-42) et des données du tableau 9-9, on obtient:

$$L(\xi_z, a_1) = \frac{p_{\vartheta_2}(z)(1-\zeta)}{\sum_{\vartheta} p_{\vartheta}(z)\,\xi(\vartheta)}; \qquad (9\text{-}44)$$

$$L(\xi_z, a_2) = \frac{w p_{\vartheta_1}(z)\,\zeta}{\sum_{\vartheta} p_{\vartheta}(z)\,\xi(\vartheta)}. \qquad (9\text{-}45)$$

La frontière z_0 correspond à l'égalité des pertes moyennes pour les décisions a_1 a a_2, ce qui donne:

$$(1-\zeta)\, p_{\vartheta_2}(z_0) = \zeta w p_{\vartheta_1}(z_0) \qquad (9\text{-}46)$$

ou

$$\frac{p_{\vartheta_2}(z_0)}{p_{\vartheta_1}(z_0)} = \frac{\zeta w}{1-\zeta} = c(\zeta, w). \tag{9-47}$$

On voit qu'à chaque valeur de ζ va correspondre sa frontière z_0 et par conséquent les domaines correspondants S_ζ et $C(S_\zeta)$. En même temps, la grandeur ζ va définir également les probabilités des décisions erronées $\alpha(\zeta)$ et $\beta(\zeta)$ qui représentent, pour $\vartheta = \vartheta_1$, les probabilités pour que le point z tombe dans le domaine $C(S_\zeta)$ et que soit prise la décision a_2 et, pour $\vartheta = \vartheta_2$, la probabilité pour que le point z tombe dans le domaine S_ζ et que soit prise la décision a_1. Ces probabilités sont définies par les relations

$$\alpha(\zeta) = P(a_2 \mid \vartheta_1) = \sum_{z \in C(S_\zeta)} p_{\vartheta_1}(z); \tag{9-48}$$

$$\beta(\zeta) = P(a_1 \mid \vartheta_2) = \sum_{z \in S_\zeta} p_{\vartheta_2}(z) \tag{9-49}$$

Pour voir de quelle façon les probabilités de décisions erronées dépendent de ζ, trouvons les valeurs de $\alpha(\zeta)$ et de $\beta(\zeta)$ pour les valeurs extrêmes $\zeta = 0$ et $\zeta = 1$.

Les éléments $z \in Z$ compris dans l'ensemble S sont définis par la condition

$$L(\xi_z, a_1) \leqslant L(\xi_z, a_2)$$

ou

$$\frac{p_{\vartheta_2}(z)}{p_{\vartheta_1}(z)} \leqslant \frac{\zeta w}{1-\zeta}. \tag{9-50}$$

Pour $\zeta = 0$, cette condition donne:

$$\frac{p_{\vartheta_2}(z)}{p_{\vartheta_1}(z)} = 0,$$

ce qui est possible seulement si $S_\zeta = \varnothing$, $C(S_\zeta) = Z$. Dans ce cas, (9-48) et (9-49) donnent $\alpha(0) = 1$, $\beta(0) = 0$.

Pour $\zeta = 1$, la condition (9-50) donne:

$$\frac{p_{\vartheta_2}(z)}{p_{\vartheta_1}(z)} \leqslant \infty$$

ce qui définit le domaine $S_\zeta = Z$, $C(S_\zeta) = \varnothing$, de sorte que $\alpha(1) = 0$, $\beta(1) = 1$.

Ainsi, lorsque ζ varie de 0 à 1, $\alpha(\zeta)$ varie de 1 à 0 et $\beta(\zeta)$ varie de 0 à 1.

Pour déterminer les pertes moyennes pour toutes les valeurs possibles de ζ, trouvons le risque de Bayes $\rho^*(\zeta)$ pour la fonction de décision $d(z)$ considérée. En utilisant l'expression (9-29) pour la

fonction de décision $\rho(\vartheta, d)$ et en prenant la moyenne de cette fonction suivant tous les états de la nature, on trouve:

$$\rho^*(\zeta) = \zeta w \alpha(\zeta) + (1 - \zeta)\beta(\zeta). \tag{9-51}$$

Le graphique de la dépendance $\rho^*(\zeta)$, construit d'après cette formule compte tenu du caractère trouvé de la variation de $\alpha(\zeta)$ et de $\beta(\zeta)$, est donné sur la figure 9-8. Ce graphique montre que $\rho^*(\zeta)$ s'annule pour $\zeta = 0$ et pour $\zeta = 1$, et atteint son maximum pour une

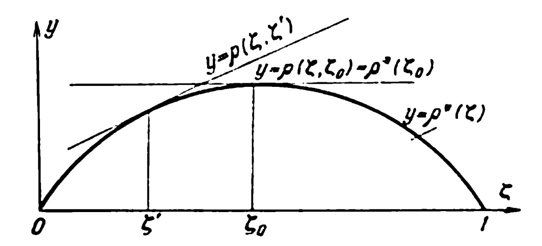


Fig. 9-8. Courbe du risque de Bayes dans un problème bialternatif

certaine valeur $\zeta = \zeta_0$. La valeur ζ_0 définit la stratégie de la nature la plus défavorable pour le statisticien qui, en adoptant le principe de Bayes, subit les pertes $\rho^*(\zeta_0)$.

Pratiquement, on rencontre très souvent des cas où les données statistiques préalables ne suffisent pas pour déterminer avec précision la probabilité a priori ζ. C'est pourquoi il est intéressant de trouver la valeur du risque de Bayes pour le cas où le statisticien prend comme point de départ une certaine valeur de la probabilité a priori ζ' et détermine par suite les grandeurs $S_{\zeta'}$, $\alpha(\zeta')$ et $\beta(\zeta')$, tandis que la valeur réelle de ζ sera toute autre. Dans ce cas, le risque de Bayes est défini par l'expression

$$\rho(\zeta, \zeta') = \zeta w \alpha(\zeta') + (1 - \zeta)\beta(\zeta'), \tag{9-52}$$

qui est linéaire par rapport à ζ et définit la tangente à la courbe $y = \rho^*(\zeta)$ au point $\zeta = \zeta'$. La figure 9-8 montre que les pertes définies par la fonction $\rho(\zeta, \zeta')$ seront supérieures à celles définies par la fonction $\rho^*(\zeta)$ si $\zeta \neq \zeta'$, de sorte qu'une connaissance approximative de la distribution a priori de probabilités entraîne une augmentation des pertes. Bien plus, si ζ diffère sensiblement de ζ', les pertes définies par la fonction $\rho(\zeta, \zeta')$ peuvent dépasser la valeur de $\rho^*(\zeta_0)$ correspondant aux pertes de Bayes maximales. Toutefois, si l'on part de la distribution de probabilités la plus défavorable correspondant à ζ_0, la fonction $y = \rho(\zeta, \zeta_0)$ sera constante et elle définira, pour tout ζ, des pertes égales à $\rho^*(\zeta_0)$. Ce cas correspond à l'adoption de la stratégie minimax.

19—0732

De cette façon, l'utilisation du principe de Bayes n'est rationnelle que dans les cas où la distribution a priori de probabilités $\xi(\vartheta)$ est connue avec une précision suffisante. Mais si la distribution a priori de probabilités $\xi(\vartheta)$ n'est pas connue ou bien n'est connue qu'approximativement, l'utilisation du principe du minimax peut s'avérer plus avantageuse.

9-5. JEUX STATISTIQUES À ÉCHANTILLONNAGES SÉQUENTIELS

a) Remarques préliminaires

Lors de l'examen des jeux statistiques à expérience unique, il a été noté que cette expérience peut être constituée par une séquence d'épreuves dont le volume et l'ordre doivent être déterminés à l'avance. De cette façon, si l'on établit à l'avance que la décision sera prise à la suite de N épreuves consécutives, toute cette séquence d'épreuves est considérée comme une expérience unique dont l'issue sera la grandeur multidimensionnelle $z = (z_1, \ldots, z_N)$, où z_i $(i = 1, \ldots, N)$ est l'issue de l'i-ième épreuve. Pour cette raison, le jeu à expérience unique est souvent appelé jeu à volume d'échantillonnage donné, en entendant par le volume d'échantillonnage le nombre d'épreuves consécutives effectuées.

Dans les jeux à volume d'échantillonnage donné, le statisticien ne peut utiliser que les stratégies qui déterminent le caractère de ses actions après l'achèvement complet de toute la séquence d'épreuves. Toutefois, le statisticien peut agir d'une autre façon. Ainsi, au lieu d'effectuer toutes les N épreuves à la fin de chaque épreuve il peut décider d'arrêter l'expérience et de prendre une décision contenue dans l'ensemble A sur la base de l'information qu'il possède déjà, ou bien il peut décider de procéder à l'épreuve suivante. Cela élargit la classe de stratégies possibles du statisticien, car au choix de la décision contenue dans l'ensemble A s'ajoute le choix de la décision d'arrêter ou de continuer l'expérience. Ces jeux sont appelés *jeux à échantillonnage séquentiel*. Avec cela, si l'on fixe le nombre limite admissible d'épreuves à la suite desquelles une décision comprise dans l'ensemble A doit obligatoirement être prise, on est en présence d'un *jeu à échantillonnage séquentiel tronqué*. Dans la suite, nous nous bornerons aux jeux de ce genre en renvoyant le lecteur qui s'intéresse à l'étude plus détaillée de ce problème aux ouvrages spécialisés [54, 59, 60]. Les résultats isolés successifs de l'expérience seront nommés *observations*.

Si l'expérience n'entraînait pas de frais, l'élargissement de la classe de stratégies par l'introduction des échantillonnages séquentiels ne serait pas privé de sens, car le statisticien ne pourrait que gagner en obtenant toutes les N observations. Pourtant, dans beau-

coup de cas, l'expérience coûte cher et nécessite du temps pour être faite. Dans ces conditions, le statisticien peut obtenir un gain important si à chaque stade de l'expérience il compare le coût de la continuation de l'expérience avec le gain espéré déterminé par l'information supplémentaire obtenue.

Voyons maintenant comment peut-on décrire un jeu à échantillonnage séquentiel tronqué. Désignons par Z_i l'ensemble des issues de l'i-ième épreuve. Alors, l'espace entier des issues Z de l'expérience obtenues après l'exécution de toutes les N épreuves peut s'écrire

$$Z = Z_1 \times \ldots \times Z_N. \qquad (9\text{-}53)$$

Dans un jeu à échantillonnage séquentiel, nous admettons que la décision puisse être prise non pas sur la base de toutes les N observations $z_1, \ldots, z_N$, mais en obtenant seulement les j premières observations $z_1, \ldots, z_j$. De cette façon, l'ensemble Z peut être décomposé en les sous-ensembles disjoints $S_0, S_1, \ldots, S_N$ tels que, si $z \in S_j$, la décision soit prise sur la base des j premières observations.

L'ensemble $S = (S_0, S_1, \ldots, S_N)$ est appelé *plan de l'échantillonnage séquentiel*. L'ensemble Z peut être décomposé en les sous-ensembles disjoints S_j par plusieurs procédés différents. Chacun de ces procédés va déterminer son plan de l'échantillonnage séquentiel. Désignons par $\mathfrak{S}$ l'ensemble des plans possibles de l'échantillonnage séquentiel et appelons-le *classe complète d'échantillonnages séquentiels*.

Pour avoir la possibilité de prendre une décision, il faut définir la fonction de décision $d(z)$ qui définit pour chaque séquence d'observations la décision de l'ensemble A. Tout comme dans le cas d'un jeu à expérience unique, le statisticien choisit la fonction de décision $d(z)$ dans l'espace des fonctions de décision D qui contient toutes les fonctions de décision possibles.

Dans un jeu à échantillonnages séquentiels, la stratégie du statisticien consiste tout d'abord à choisir le plan de l'échantillonnage séquentiel $S \in \mathfrak{S}$ qui indique le moment où doit prendre fin l'expérience, ensuite, à choisir la fonction de décision $d \in D$ qui détermine la décision à prendre à la fin de l'expérience. Ainsi, le couple (S, d) définit la stratégie du statisticien. Compte tenu du fait que le couple (S, d) est un élément du produit direct $\mathfrak{S} \times D$, on en conclut que $\mathfrak{S} \times D$ est l'espace des stratégies pures du statisticien.

Le nombre total de stratégies dans les jeux à échantillonnages séquentiels est beaucoup plus important que dans les jeux à expérience unique, ce qui rend difficile l'élaboration de la liste complète des stratégies du statisticien et le choix de la meilleure d'entre elles. Ce travail peut être sensiblement simplifié en calculant à chaque stade de l'expérience la distribution a posteriori de probabilités sur la base de toute l'information accumulée à ce moment. Au fur et à mesure de l'obtention des observations consécutives

l'indétermination relative à l'état réel de la nature diminue et l'occasion se présente d'élaborer les critères de détermination du moment où l'expérience doit être achevée pour prendre une décision de l'ensemble A.

b) Utilisation de la distribution a posteriori de probabilités pour la détermination des règles de Bayes séquentielles

Nous n'allons considérer que des jeux statistiques dont les résultats des épreuves isolées $z_1, \ldots, z_N$ sont des variables aléatoires indépendantes. Désignons par $q_\vartheta(z_j)$ la distribution de probabilités sur l'espace Z_j pour un état de la nature donné ϑ. Le coût d'une épreuve isolée sera considéré constant $c = 1$.

Pour décrire le plan des échantillonnages séquentiels, prenons en qualité de base la distribution de probabilités $\xi(\vartheta)$ sur l'espace des états de la nature Θ, en entendant par la distribution $\xi(\vartheta)$ non seulement la distribution a priori de probabilités précédant l'expérience, mais aussi la distribution *a posteriori* de probabilités après l'accomplissement d'un certain nombre d'épreuves. Dans ce qui suit, on va désigner par $\xi_0(\vartheta)$ la distribution a priori de probabilités, et par $\xi_j(\vartheta)$ la distribution a posteriori de probabilités après j épreuves.

La distribution de probabilités $\xi_j(\vartheta)$ comprendra toute l'information concernant l'état de la nature ϑ aussi bien celle qui précédait l'expérience, que celle obtenue à la suite des j premières observations. Pour cette raison, la distribution $\xi_j(\vartheta)$ peut être considérée comme étant la distribution a priori précédant la $(j+1)$-ième observation. Il s'ensuit que la distribution $\xi_{j+1}(\vartheta)$ est exprimée en fonction de $\xi_j(\vartheta)$ à l'aide de la formule pour la distribution a posteriori de probabilités. Considérons le passage de la distribution a priori de probabilités $\xi_j(\vartheta)$ à la distribution a posteriori $\xi_{j+1}(\vartheta)$ comme une certaine transformation T de la distribution $\xi_j(\vartheta)$. Alors, la distribution a posteriori de probabilités peut s'écrire en vertu de (9-39) de la façon suivante:

$$T\xi_j(\vartheta) = \xi_{j+1}(\vartheta) = \frac{\xi_j(\vartheta)\, q_\vartheta(z_{j+1})}{\sum_\vartheta \xi_j(\vartheta)\, q_\vartheta(z_{j+1})}. \qquad (9\text{-}54)$$

Examinons le principe d'obtention du plan de l'échantillonnage séquentiel sur la base de l'utilisation de la distribution a posteriori de probabilités en prenant comme exemple un problème bialternatif, où $\Theta = \{\vartheta_1, \vartheta_2\}$ et $A = \{a_1, a_2\}$.

Désignons par $(\zeta, 1 - \zeta)$ la distribution a posteriori de probabilités sur l'espace Θ après un certain nombre d'épreuves. Dans ce problème, l'espace E des stratégies mixtes de la nature est défini par le domaine des valeurs possibles de ζ, c.-à-d. par l'intervalle $[0, 1]$ de l'axe réel.

Il se peut que $\zeta = 1$. Dans ce cas, on ne peut prendre que la décision a_1. Si $\zeta = 0$, la décision a_2 est obligatoire. Mais si, par exemple, $\zeta = 0{,}5$, aucune des décisions n'a la priorité et l'expérience doit être continuée pour préciser l'état réel de la nature.

Ici, on a examiné les cas extrêmes de la distribution de probabilités $\xi(\vartheta)$. Dans le cas général, on peut se donner les quantités δ et γ $(0 \leqslant \delta \leqslant 1,\ 0 \leqslant \gamma \leqslant 1,\ \delta \geqslant \gamma)$ telles que:

si ζ se trouve dans la gamme $[\delta, 1]$, on prend la décision a_1;

si ζ se trouve dans la gamme $[0, \gamma]$, on prend la décision a_2;

si ζ se trouve dans la gamme (γ, δ), on prend la décision d'effectuer l'épreuve suivante. Les gammes $\Delta(a_1) = [\delta, 1]$ et $\Delta(a_2) = [0, \gamma]$ sont appelées *domaines d'arrêt*. La représentation graphique de la règle qui vient d'être formulée est donnée sur la figure 9-9.

Fig. 9-9. Règle de la prise de décision dans un problème bialternatif

Les domaines d'arrêt peuvent être également déterminés pour le cas où le nombre d'état de la nature est supérieur à deux. Il est vrai que, dans ce cas, l'espace E des stratégies mixtes de la nature sera d'une forme plus compliquée. Ainsi, pour trois états de la nature, cet espace a la forme d'un triangle équilatéral dont la hauteur est égale à l'unité (cf. fig. 9-11).

En général, on appelle domaine d'arrêt $\Delta(a)$ dans l'espace E un sous-ensemble de cet espace

$$\Delta(a) \subset \mathrm{E} \tag{9-55}$$

tel que, si à la suite d'une certaine épreuve on a $\xi(\vartheta) \in \Delta(a)$, on arrête l'expérience et l'on prend la décision a.

Etant donné que chaque épreuve a son coût, il ne sera pas indifférent si $\xi(\vartheta)$ tombe dans le domaine $\Delta(a)$ avant l'expérience ou après un certain nombre d'épreuves. Cela signifie que chaque nouvelle épreuve fait varier aussi bien la valeur de $\xi(\vartheta)$ que celle de $\Delta(a)$. En outre, à chaque stade de l'expérience on ne s'intéresse pas aux épreuves déjà effectuées et dont les résultats ont été déjà utilisés, mais à celles qui doivent être faites pour parachever l'expérience. Conformément à ce qui vient d'être dit, les domaines d'arrêt au stade de l'expérience où l'on a effectué j épreuves sur N seront désignés par $\Delta_{N-j}(a)$.

Si l'espace des décisions comporte m éléments $a_1, \ldots, a_m$, à chaque stade de l'expérience il faut déterminer m domaines $\Delta(a)$ correspondant à chaque $a \in A$. Ces domaines doivent être convexes et disjoints [54].

c) Règle des échantillonnages séquentiels

Supposons que dans l'espace E soient délimités m domaines $\Delta(a)$ à chaque stade de l'expérience. Dans ce cas, la règle des échantillonnages séquentiels consistera dans ce qui suit.

On connaît à l'avance la distribution a priori de probabilités $\xi_0(\vartheta)$. Si $\xi_0(\vartheta) \in \Delta_N(a_i)$ pour un certain i, on prend la décision a_i sans faire d'expérience. Si $\xi_0(\vartheta) \notin \Delta_N(a_i)$ pour aucun i, on exécute la première épreuve et l'on calcule la distribution a posteriori de probabilités $\xi_1(\vartheta)$. Ensuite, on examine les domaines $\Delta_{N-1}(a)$. Si $\xi_1(\vartheta) \in \Delta_{N-1}(a_i)$ pour un certain i, l'expérience est arrêtée et l'on prend la décision a_i. Si $\xi_1(\vartheta) \notin \Delta_{N-1}(a_i)$ pour aucun i, on effectue l'épreuve suivante et ainsi de suite.

Généralement parlant, si l'expérience n'a pas été arrêtée après les $j-1$ premières observations, on effectue une observation supplémentaire et l'on calcule $\xi_j(\vartheta)$. Si $\xi_j(\vartheta) \in \Delta_{N-j}(a_i)$ pour un certain i, l'expérience est arrêtée et l'on prend la décision a_i. Si la décision n'a pas été prise après la $(N-1)$-ième observation, on effectue la N-ième observation et l'on fait le choix définitif. Pour satisfaire à cette condition, les domaines $\Delta_0(a_i)$ doivent être choisis de façon que

$$\bigcup_i \Delta_0(a_i) = \mathrm{E}. \qquad (9\text{-}56)$$

On voit de cette façon que le problème de la description des règles séquentielles de Bayes se ramène au problème de la définition des domaines $\Delta_{N-j}(a)$ dans l'espace E pour tous les $a \in A$ et tous les j de 0 à N.

d) Fonction de risque pour la règle séquentielle optimale

En présence d'échantillonnages séquentiels, la fonction de risque doit définir à chaque stade de l'expérience (par exemple, après j épreuves) les pertes moyennes minimales que le statisticien va subir en prenant la meilleure des décisions possibles, y compris la décision concernant la continuation de l'expérience. La prise de décision étant basée sur la connaissance de la distribution a posteriori de probabilités $\xi_j(\vartheta)$, la fonction de risque sera également fonction de la même distribution de probabilités. La fonction de risque obtenue après j épreuves sera désignée par $\rho^*(\xi_i)$.

Les expressions de la fonction de risque seront recherchées successivement en partant du dernier stade de l'expérience. Supposons que l'on ait effectué toutes les N épreuves et que l'on ait trouvé la distribution a posteriori de probabilités $\xi_N(\vartheta)$. Dans ce cas, les pertes moyennes subies par le statisticien qui prend la décision

$a \in A$ sont égales à

$$L(\xi_N, a) = \sum_{\vartheta} L(\vartheta, a)\, \xi_N(\vartheta). \qquad (9\text{-}57)$$

La valeur minimale de ces pertes correspondant à la décision de Bayes a^* est donnée par l'expression

$$R^*(\xi_N) = L(\xi_N, a^*) = \min_a L(\xi_N, a). \qquad (9\text{-}58)$$

Etant donné qu'il n'est pas possible de prendre une décision meilleure que a^*, la grandeur $R^*(\xi_N)$ coïncidera à ce stade de l'expérience avec la fonction de risque

$$\rho^*(\xi_N) = R^*(\xi_N). \qquad (9\text{-}59)$$

Considérons maintenant un stade arbitraire de l'expérience et trouvons les pertes moyennes minimales subies par le statisticien après j épreuves ($j = 0, 1, \ldots, N - 1$). Avec cela, le statisticien dispose de la distribution de probabilités $\xi_j(\vartheta)$.

Dans ce cas, le statisticien peut choisir l'une des deux alternatives suivantes:

1. Il arrête l'expérience et prend la décision $a \in A$. Ainsi, les pertes moyennes minimales qu'il subit sont

$$R^*(\xi_j) = \min_a L(\xi_j, a) = \min_a \sum_{\vartheta} L(\vartheta, a)\, \xi_j(\vartheta). \qquad (9\text{-}60)$$

2. Il peut prendre la décision d'effectuer la $(j + 1)$-ième épreuve. Dans ce cas, il subira des pertes égales à la somme du coût d'une épreuve (égal à l'unité) et des pertes après la prise de la meilleure décision d'après les résultats de la $(j + 1)$-ième observation, c.-à-d. $\rho^*(\xi_{j+1})$.

Mais le statisticien ne connaît pas la distribution de probabilités $\xi_{j+1}(\vartheta)$. Il ne connaît que la distribution $\xi_i(\vartheta)$ et ne peut déterminer $\xi_{j+1}(\vartheta)$ que d'après la formule (9-54) en tant que $T\xi_j(\vartheta)$, grandeur qui peut être trouvée pour des valeurs concrètes de $z_{j+1} \in Z_{j+1}$ et $\vartheta \in \Theta$ que le statisticien ignore pour le moment. C'est pourquoi il ne doit pas avoir en vue une valeur concrète de la fonction de risque $\rho^*(\xi_{j+1})$, mais seulement la valeur moyenne de $M[\rho^*(T\xi_j)]$, la médiation devant être réalisée suivant toutes les valeurs possibles de $z_{j+1} \in Z_{j+1}$ avec la distribution de probabilités $q_\vartheta(z_{j+1})$ et suivant tous les $\vartheta \in \Theta$ avec la distribution de probabilités $\xi_j(\vartheta)$:

$$M[\rho^*(T\xi_j)] = \sum_{\vartheta,\, z_{j+1}} \rho^*[T\xi_j(\vartheta)]\, q_\vartheta(z_{j+1})\, \xi_j(\vartheta). \qquad (9\text{-}61)$$

Ainsi, les pertes minimales du statisticien qui décide de continuer l'expérience sont:

$$1 + M[\rho^*(T\xi_j)]. \qquad (9\text{-}62)$$

Après j épreuves, la fonction de risque sera définie par le minimum des pertes moyennes en examinant les deux façons d'agir du statisticien : arrêt ou continuation de l'expérience. Donc,

$$\rho^*(\xi_j) = \min\{R^*(\xi_j),\ 1 + M[\rho^*(T\xi_j)]\}. \qquad (9\text{-}63)$$

L'expression (9-63) peut être présentée sous une forme plus condensée sans utiliser l'indice j désignant le numéro de l'épreuve. Considérons le stade de l'expérience où jusqu'à la fin de cette dernière il nous reste encore k épreuves. Désignons par $\xi(\vartheta)$ la distribution a posteriori de probabilités à ce stade et par $\xi'(\vartheta) = T\xi(\vartheta)$ la distribution a posteriori de probabilités après avoir effectué encore une épreuve. Désignons par $\rho_k^*(\xi)$ la fonction de risque au stade où jusqu'à la fin de l'expérience il reste encore k épreuves. Alors les expressions (9-59) et (9-63) s'écriront :

$$\rho_0^*(\xi) = R^*(\xi); \qquad (9\text{-}64)$$

$$\rho_k^*(\xi) = \min\{R^*(\xi),\ 1 + M[\rho_{k-1}^*(T\xi)]\}. \qquad (9\text{-}65)$$

Désignons par V l'espace des issues de l'épreuve suivante dont les éléments seront notés v. En ces notations, on aura :

$$R^*(\xi) = \min_a \sum_\vartheta L(\vartheta, a)\,\xi(\vartheta); \qquad (9\text{-}66)$$

$$M[\rho_{k-1}^*(T\xi)] = \sum_{\vartheta, v} \rho_{k-1}^*[T\xi(\vartheta)]\,q_\vartheta(v)\,\xi(\vartheta); \qquad (9\text{-}67)$$

$$T\xi(\vartheta) = \frac{\xi(\vartheta)\,q_\vartheta(v)}{\sum_\vartheta \xi(\vartheta)\,q_\vartheta(v)}. \qquad (9\text{-}68)$$

Les formules (9-64) et (9-65) peuvent être utilisées pour définir les domaines d'arrêt. Mais des résultats relativement simples sont obtenus seulement pour le problème bialternatif auquel nous allons borner notre exposé.

e) Détermination des domaines d'arrêt pour un problème bialternatif en présence d'un échantillonnage séquentiel tronqué

Examinons le cas d'un problème bialternatif où la fonction de pertes est définie par la matrice

$$Q = \begin{Vmatrix} 0 & q_{12} \\ q_{21} & 0 \end{Vmatrix}, \qquad (9\text{-}69)$$

la distribution de probabilités sur l'espace des états de la nature étant $\xi(\vartheta) = (\zeta, 1 - \zeta)$. Désignons par $\Delta_k(a_1)$ et $\Delta_k(a_2)$ les domaines d'arrêt correspondant aux décisions a_1 et a_2 au moment où jusqu'à

la fin de l'expérience il reste encore à effectuer k observations. Ces domaines sont définis par des valeurs $\zeta = \delta_k$ et $\zeta = \gamma_k$, $\gamma_k \leqslant \delta_k$, telles que

$$\Delta_k (a_1) = [\delta_k \leqslant \zeta \leqslant 1], \quad \Delta_k (a_2) = [0 \leqslant \zeta \leqslant \gamma_k]. \qquad (9\text{-}70)$$

La définition des domaines d'arrêt se réduit à la détermination des valeurs de γ_k et δ_k pour $k = 0, 1, \ldots, N$. On y parvient à l'aide des expressions obtenues pour la fonction de risque.

Commençons la détermination des points frontières des domaines d'arrêt à partir de $k = 0$. En exprimant $\xi (\vartheta)$ en fonction de ζ, il vient de (9-64) et (9-66):

$$\rho_0^* (\zeta) = R^* (\zeta) = \min [(1 - \zeta) q_{21}, \zeta q_{12}]. \qquad (9\text{-}71)$$

Conformément à (9-56), $\Delta_0 (a_1)$ et $\Delta_0 (a_2)$ doivent être des ensembles réels disjoints dont la réunion donne le segment fermé 0,1. Il s'ensuit que ces domaines doivent être séparés l'un de l'autre par le point $\zeta = \gamma_0 = \delta_0$ qui est déterminé par la condition

$$(1 - \zeta) q_{21} = \zeta q_{12}. \qquad (9\text{-}72)$$

La condition (9-72) est la condition d'égalité des pertes moyennes minimales lors de la prise des décisions a_1 et a_2. A partir de (9-72) on trouve:

$$\gamma_0 = \delta_0 = \frac{q_{21}}{q_{12} + q_{21}}. \qquad (9\text{-}73)$$

Pour $k \neq 0$ arbitraire, la fonction de risque est de la forme

$$\rho_k^* (\zeta) = \min \{R^* (\zeta), 1 + M [\rho_{k-1}^* (T\zeta)]\}. \qquad (9\text{-}74)$$

Les conditions d'obtention des valeurs frontières de ζ pour les ensembles $\Delta_k (a_1)$ et $\Delta_k (a_2)$ seront des conditions d'égalité des pertes moyennes minimales lors de la prise des décisions a_1 ou a_2 et de la décision de continuer l'expérience, ce qui s'exprime par la relation

$$R^* (\zeta) = 1 + M [\rho_{k-1}^* (T\zeta)]. \qquad (9\text{-}75)$$

En observant que, pour R^* (ζ) de l'expression (9-71), la grandeur ζq_{21} est une fonction monotonement croissante de ζ qui s'annule pour $\zeta = 0$, tandis que $(1 - \zeta)$ q_{12} est une fonction monotonement décroissante de ζ, qui s'annule pour $\zeta = 1$, et en substituant à ζ, aux points frontières, γ_k pour le domaine $\Delta_k (a_2)$ et δ_k pour le domaine $\Delta_k (a_1)$, on peut mettre la condition (9-75) sous la forme

$$\gamma_k q_{21} = 1 + M [\rho_{k-1}^* (T\gamma_k)]; \qquad (9\text{-}76)$$

$$(1 - \delta_k) q_{12} = 1 + M [\rho_{k-1}^* (T\delta_k)]. \qquad (9\text{-}77)$$

La représentation graphique de ces conditions est donnée sur la figure 9-10.

Les relations obtenues montrent que le problème de définition de γ_k et δ_k ainsi que de la fonction de risque $\rho_k^*(\zeta)$ se ramène au calcul de la grandeur $M[\rho_{k-1}^*(T\zeta)]$. La formule (9-67) donne l'expression générale de cette grandeur pour Θ arbitraire. Pour un problème bialternatif, cette expression prend la forme

$$M[\rho_{k-1}^*(T\zeta)] = \zeta \sum_v \rho_{k-1}^*(T\zeta)\, q_{\vartheta_1}(v) + \\ + (1-\zeta) \sum_v \rho_{k-1}^*[T(1-\zeta)]\, q_{\vartheta_2}(v), \quad (9\text{-}78)$$

où

$$T\zeta = \frac{\zeta q_{\vartheta_1}(v)}{\zeta q_{\vartheta_1}(v) + (1-\zeta)\, q_{\vartheta_2}(v)}; \quad (9\text{-}79)$$

$$T(1-\zeta) = \frac{(1-\zeta)\, q_{\vartheta_2}(v)}{\zeta q_{\vartheta_1}(v) + (1-\zeta)\, q_{\vartheta_2}(v)}. \quad (9\text{-}80)$$

Pour $k = 1$, la fonction $\rho_0^*(\zeta)$ définie par l'expression (9-71) est une fonction très simple de ζ, de sorte que la grandeur $M[\rho_0^*(T\zeta)]$

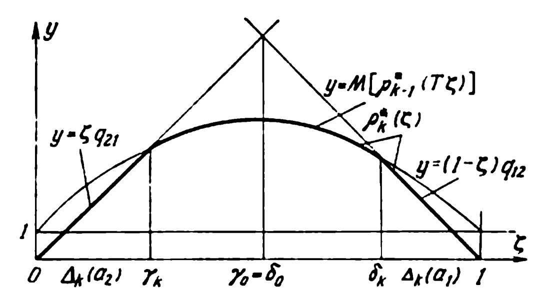


Fig. 9-10. Détermination des domaines d'arrêt dans un problème bialternatif

peut être déterminée directement à partir de (9-78). Si l'on connaît $M[\rho_0^*(T\zeta)]$, on peut calculer $\rho_1^*(\zeta)$, d'après (9-74), donc, déterminer $M[\rho_1^*(T\zeta)]$ d'après (9-78), ensuite $\rho_2^*(\zeta)$ d'après (9-74), etc. Bien que ce procédé de recherche de $\rho_k^*(\zeta)$ nécessite des calculs importants, il limite les opérations plus compliquées à la détermination des espérances mathématiques.

PROBLÈMES AU CHAPITRE 9

9-1. D'après les données du tableau 9-3, trouver la stratégie basée sur le principe du minimax des pertes supplémentaires dans le problème de la ligne technologique.

9-2. En utilisant la construction géométrique de la figure 9-6, effectuer les calculs nécessaires pour la détermination des stratégies minimax et de Bayes dans le problème de la ligne technologique à expérience unique.

9-3. Un signal u transmis par une voie de communication est entaché de parasites s distribués suivant la loi normale $N(0, \sigma^2)$ de façon qu'à la réception on mesure la grandeur $z = u + s$. Le signal u peut prendre deux valeurs : $u = 0$ qui veut dire qu'il n'y a pas de signal (ϑ_1), et $u = u_0 = \text{const}$ qui signifie que

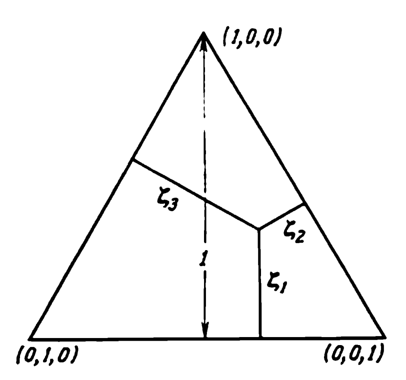


Fig. 9-11. Espace des stratégies mixtes de la nature pour le cas où la nature peut se trouver dans un des trois états

le signal est présent (ϑ_2). Le seuil du récepteur est z_0. Ce récepteur émet la décision que le signal u_0 est transmis si $z > z_0$. Trouver la valeur du seuil z_0 en utilisant les méthodes de résolution d'un problème bialternatif et en posant $\zeta = 0{,}25$ et $w = 2$.

9-4. Montrer que, si la nature peut se trouver dans un des trois états $\Theta = \{\vartheta_1, \vartheta_2, \vartheta_3\}$, toutes les stratégies mixtes possibles du statisticien $\xi(\vartheta) = (\zeta_1, \zeta_2, \zeta_3)$ forment un espace des stratégies mixtes en forme de triangle dont la hauteur est égale à l'unité, comme on le voit sur la figure 9-11.

CHAPITRE 10

PROGRAMMATION DYNAMIQUE

10-1. COMMANDE OPTIMALE COMME PROBLÈME VARIATIONNEL

a) Formulation mathématique du problème de commande optimale

La construction des systèmes de commande automatique modernes est caractérisée par la tendance d'obtenir des systèmes qui, dans un certain sens, soient optimaux. Lorsqu'il s'agit de la commande des processus technologiques, cette tendance trouve son expression en l'obtention d'une quantité maximale de produits de haute qualité, l'utilisation des ressources (en matières premières, énergétiques, etc.) restant limitée. Dans les systèmes de commande des navires, des avions ou des fusées, on tâche de minimiser le temps nécessaire pour que l'objet commandé arrive à un point donné ou soit placé sur la trajectoire nécessaire, en limitant en même temps l'angle de braquage des gouvernails, la consommation de combustible, etc. Dans les systèmes asservis et de stabilisation, l'attention est portée sur l'obtention du maximum de précision en présence de toute sorte de contraintes imposées aux coordonnées de l'objet à régler, aux organes d'exécution et au régulateur. Dans tous les exemples susmentionnés, les problèmes de commande se ramènent à la découverte du meilleur (dans un certain sens) processus parmi une multitude de processus possibles, donc ils se rapportent à la classe de problèmes de commande dynamiques.

Au chapitre 6 on a vu que la formulation des problèmes dynamiques de commande optimale se réduit à ce qui suit. On a un objet commandé dont l'état est caractérisé par une variable multidimensionnelle $x = (x^{(1)}, \ldots, x^{(N)})$. Le caractère des processus ayant lieu dans l'objet commandé peut être changé en mettant en œuvre telle ou telle commande u de l'espace des commandes admissibles U. En général, la commande $u \in U$ peut aussi être une grandeur multidimensionnelle $u = (u^{(1)}, \ldots, u^{(R)})$.

Le caractère de l'évolution de l'objet commandé est décrit par un système d'équations différentielles

$$\dot{x} = g(x, u), \quad x(0) = c. \tag{10-1}$$

En tant que critère de qualité de la commande, on prend une estimation intégrale de la forme

$$J_1(u) = \int_0^T Q_1[x(t), u(t)]\, dt \tag{10-2}$$

ayant le sens physique de pertes, où T est le temps de déroulement du processus de commande et $Q_1[x(t), u(t)] = q_1(t)$ sont les pertes instantanées au moment t pour l'état $x(t)$ du système et la commande $u(t)$. En qualité de contraintes additionnelles on peut avoir des contraintes imposées aux ressources ou aux limites de variation de certains paramètres et qui s'expriment mathématiquement par la relation

$$\int_0^T H[x(t), u(t)]\, dt \leqslant K, \qquad (10\text{-}3)$$

coïncidant avec (6-33).

Au chapitre 6 il a été indiqué qu'une commande u de l'ensemble des commandes admissibles U est appelée commande *optimale* lorsque, pour l'objet décrit par l'équation différentielle (10-1), en présence des contraintes données imposées aux ressources (10-3), le critère de qualité de la commande (10-2) atteint sa valeur minimale (maximale).

Le problème de la commande optimale formulé de cette façon se rapporte à la classe de problèmes variationnels dont l'étude fait l'objet de la branche des mathématiques appelée *calcul variationnel*. La grandeur $J_1(u)$ définie par la relation (10-2) est appelée *fonctionnelle*. A la différence d'une fonction, par exemple $f(x)$, dont les valeurs numériques sont définies sur l'ensemble des valeurs de l'argument x, les valeurs numériques de la fonctionnelle $J_1(u)$ sont définies sur l'ensemble de toutes les commandes possibles $u(t)$. Le problème de recherche de la commande optimale se ramène à choisir de l'ensemble des commandes admissibles $u(t)$ une commande telle que la fonctionnelle $J_1(u)$ prenne sa valeur numérique minimale.

D'habitude, aux problèmes nécessitant la minimisation d'une fonctionnelle de la forme (10-2) subordonnée à la relation différentielle (10-1), en présence de la contrainte intégrale (10-3), on substitue la minimisation d'une nouvelle fonctionnelle

$$J(u) = \int_0^T Q_1(x, u)\, dt + \lambda \int_0^T H(x, u)\, dt, \qquad (10\text{-}4)$$

subordonnée exclusivement à la relation différentielle (10-1). Le paramètre λ de la fonctionnelle (10-4), appelé multiplicateur de Lagrange, joue dans les problèmes de commande optimale le rôle de « prix » des ressources limitées. Sa valeur est trouvée à partir des conditions aux limites du problème variationnel.

La possibilité de simplifier un problème variationnel à contraintes intégrales par l'introduction des multiplicateurs de Lagrange est basée sur le théorème ci-après.

Théorème 10-1. *Si u (t) est la commande optimale, pour laquelle la fonctionnelle* (10-4) *atteint son minimum absolu, et si la contrainte* (10-3) *est respectée, alors pour u (t) est atteint le minimum absolu de la fonctionnelle* (10-2) *subordonnée à la contrainte* (10-3).

Démonstration. On utilise la démonstration par l'absurde. Soit $v(t)$ une autre commande différente de $u(t)$ et telle que

$$\int_0^T Q_1(x, v)\, dt < \int_0^T Q_1(x, u)\, dt \tag{10-5}$$

et que la condition

$$\int_0^T H(x, v)\, dt \leqslant K \tag{10-6}$$

soit satisfaite. Alors,

$$\int_0^T Q_1(x, v)\, dt + \lambda \int_0^T H(x, v)\, dt \leqslant \int_0^T Q_1(x, v)\, dt + \lambda K <$$
$$< \int_0^T Q_1(x, u)\, dt + \lambda K = \int_0^T Q_1(x, u)\, dt + \lambda \int_0^T H(x, u)\, dt, \tag{10-7}$$

ce qui contredit l'hypothèse que $u(t)$ minimise (10-4).

L'application des procédés du calcul variationnel au problème de recherche de la commande optimale ne s'est pas répandue à cause de toute une série de complications qui surgissent à cet effet. Pour cette raison, nous n'allons pas nous attarder sur les méthodes de résolution du problème variationnel en priant le lecteur intéressé de s'adresser aux ouvrages spécialisés [61]. Pour le moment, on ne relèvera que certains points nécessaires à l'exposé qui suit.

Une des plus importantes notions du calcul variationnel est celle de *variation d'une fonction* qui, lors de l'étude des fonctionnelles, joue le même rôle que la différentielle lors de l'étude des fonctions.

Soit $f(x)$ une fonction continue sur l'intervalle $[a, b]$. Examinons un point x intérieur à cet intervalle et une certaine valeur fixée de la différentielle de l'argument de la fonction $\Delta x = dx$. La différence

$$f(x + \Delta x) - f(x) = df(x) = f'(x)\, \Delta x \tag{10-8}$$

est appelée différentielle de la fonction $f(x)$ au point x. On sait que la condition $df(x) = 0$ est une condition nécessaire pour que la fonction $f(x)$ atteigne son minimum (maximum) au point x.

Pour aboutir à une analogie avec le calcul variationnel, il est commode de définir la différentielle $df(x)$ d'une façon un peu diffé-

rente. Examinons la valeur de la fonction $f(x + \varepsilon\Delta x)$. Pour x et Δx fixés, elle sera fonction de ε. En dérivant cette fonction par rapport à ε et en posant $\varepsilon = 0$, on obtient:

$$\frac{\partial}{\partial\varepsilon} f(x+\varepsilon\Delta x)\Big|_{\varepsilon=0} = f'(x)\,\Delta x = df(x). \tag{10-9}$$

Examinons maintenant les notions analogues du calcul variationnel. Soient $u(t)$ et $u_1(t)$ deux commandes utilisées. La différence

$$\delta u(t) = u_1(t) - u(t) \tag{10-10}$$

est appelée variation de la fonction $u(t)$, tandis que la différence

$$\delta J(u) = J(u + \delta u) - J(u) \tag{10-11}$$

est appelée *variation de la fonctionnelle.*

La variation de la fonctionnelle peut être définie d'une autre façon. A cette fin, examinons, pour $u(t)$ et $\delta u(t)$ fixées, la fonctionnelle

$$J(u + \varepsilon\delta u) = \varphi(\varepsilon) \tag{10-12}$$

qui est fonction de ε. Supposons que cette fonctionnelle soit définie pour différents ε, ce qui veut dire que les différentes commandes $u(t) + \varepsilon\, u(t)$ sont possibles au voisinage de la commande fixée $u(t)$, c.-à-d. $u(t)$ est un point intérieur de l'espace des commandes admissibles U. Dans ce cas, la variation de la fonctionnelle peut être définie par analogie à la différentielle d'une fonction, donnée par la relation (10-9), comme

$$\delta J(u) = \frac{\partial}{\partial\varepsilon} J(u+\varepsilon\delta u)\Big|_{\varepsilon=0}. \tag{10-13}$$

Si $u(t)$ est la commande optimale, la fonction $\varphi(\varepsilon)$ atteindra son minimum pour $\varepsilon = 0$. Dans ce cas,

$$\frac{\partial\varphi(\varepsilon)}{\partial\varepsilon}\Big|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon} J(u+\varepsilon\delta u)\Big|_{\varepsilon=0} = 0, \tag{10-14}$$

c.-à-d. $\delta J(u) = 0$. Donc, la commande optimale rendra nulle la variation de la fonctionnelle.

Dans le calcul variationnel, la condition $\delta J(u) = 0$ est utilisée pour obtenir l'équation différentielle dite d'Euler dans l'ensemble des solutions de laquelle on cherche ensuite la commande $u(t)$ qui rend minimale la fonctionnelle (10-4).

b) Difficultés liées à la résolution du problème variationnel

Lors de la recherche de la commande optimale à l'aide des procédés variationnels, on se heurte à toute une série de difficultés dont certaines revêtent un caractère de principe;

1) les procédés variationnels permettent de trouver seulement les maxima et les minima relatifs de la fonctionnelle $J(u)$, or ce sont le maximum et le minimum absolus qui nous intéressent ;

2) pour beaucoup de problèmes techniques, les équations d'Euler ne sont pas linéaires, ce qui rend souvent impossible l'obtention d'une solution explicite du problème variationnel;

3) d'habitude, les valeurs des signaux de commande sont soumises aux contraintes qui rendent impossible la recherche de la commande optimale par des procédés variationnels.

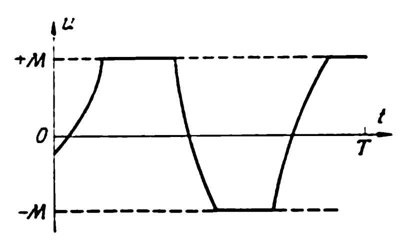


Fig. 10-1. Forme caractéristique du signal de commande optimal

Etant donné que cette dernière circonstance a eu une importance décisive pour le développement des idées nouvelles dans le domaine de la commande optimale, examinons-la plus en détail.

Les contraintes habituelles imposées aux signaux de commande sont de la forme

$$|u_i(t)| \leqslant M_i, \tag{10-15}$$

ce qui signifie que la valeur des signaux appliqués aux organes de commande ne doit pas dépasser certaines limites. Ainsi, la tension limite appliquée à l'induit d'un moteur électrique, l'angle de braquage limite du gouvernail d'un avion, la température limite dans la chambre de combustion d'un moteur à réaction, etc., ne peuvent prendre que des valeurs limitées. Avec cela, l'obtention des processus optimaux exige, en règle générale, que les signaux de commande soient maintenus à leurs valeurs limites, ce qui correspond au développement le plus rapide et le plus efficace des processus dans l'objet commandé. Sur la figure 10-1 est représentée la variation typique, pour ces cas, de la commande $u(t)$ pour un processus optimal.

Pourtant, les valeurs limites de la commande $u(t)$ se situent sur les frontières du domaine des commandes admissibles U, donc elles ne sont pas des points intérieurs à ce domaine auxquels on pourrait appliquer des procédés variationnels. Il est vrai que l'on peut se débarrasser des contraintes de la forme (10-15) en introduisant des variables nouvelles v_i liées aux variables u_i par la relation $u_i = M_i \sin v_i$. Dans ce cas, aux valeurs $|u_i| = M_i$ vont correspondre les

valeurs $v_i = \pm \pi/2$ qui sont des points intérieurs au domaine des nouvelles commandes admissibles. Mais d'habitude, ce changement de variables entraîne une complication considérable des équations obtenues.

Les difficultés auxquelles se heurte la résolution du problème de l'optimisation de la commande ont attiré l'attention de beaucoup de savants en U.R.S.S. et aux U.S.A. Le mathématicien soviétique L. Pontriaguine et ses élèves V. Boltianski, R. Gamkrélidzé et E. Michtchenko ont élaboré une théorie de la commande optimale basée sur le *principe du maximum* formulé par L. Pontriaguine. Ce principe a permis d'élaborer une base strictement mathématique de la théorie de la commande optimale, en ouvrant ainsi de larges perspectives à son application pratique dans le domaine des systèmes de commande automatiques. Etant donné que les cadres du présent livre ne nous permettent pas d'exposer la théorie générale de la commande optimale basée sur le principe du maximum, nous sommes obligés de nous référer à la bibliographie [62 à 64].

Le mathématicien américain R. Bellman a trouvé une autre voie permettant de calculer les processus optimaux connue sous le nom de *programmation dynamique* [65 à 67]. La méthode de programmation dynamique fournit à l'ingénieur une procédure efficace de résolution du problème d'optimisation de la commande, qui se prête bien à l'utilisation des calculatrices numériques. Dans ce qui suit, cette méthode sera examinée plus en détail.

10-2. MÉTHODE DE PROGRAMMATION DYNAMIQUE

a) Forme discrète du problème variationnel

Pour surmonter les difficultés rencontrées lors de la résolution d'un problème variationnel, on fait appel aux méthodes de calcul efficaces comme, par exemple, la méthode de programmation dynamique. Cette méthode permet de trouver la commande optimale dans les problèmes à étapes multiples. Elle peut être aussi appliquée à la résolution des problèmes variationnels si ceux-ci sont présentés sous forme discrète.

En utilisant le théorème 10-1, formulons le problème variationnel de la façon suivante: pour l'objet décrit par l'équation différentielle (10-1), trouver, dans le domaine des commandes admissibles U, la commande $u(t)$ qui minimise la fonctionnelle

$$J(u) = \int_0^T Q[x(t), u(t)]\, dt, \tag{10-16}$$

où

$$Q(x, u) = Q_1(x, u) + \lambda H(x, u). \tag{10-17}$$

La forme discrète de ce problème sera obtenue si l'on effectue le choix de la commande $u(t)$ seulement aux moments discrets $t = k\delta$, $k = 0, 1, \ldots, n-1$, où $\delta = T/n$. Avec cela, au lieu des fonctions $x(t)$ et $u(t)$ nous allons examiner les suites

$$x_k = x(t)\,|_{t=k\delta}, \quad u_k = u(t)\,|_{t=k\delta}.$$

En substituant dans l'équation (10-1) le rapport des accroissements $(x_{k+1} - x_k)/\delta$ à la dérivée $\dot{x} = dx/dt$, on obtient à la place de l'équation différentielle (10-1) l'équation aux différences finies

$$\frac{x_{k+1} - x_k}{\delta} = g(x_k, u_k). \tag{10-18}$$

D'habitude, cette équation s'écrit sous une forme plus commode et est résolue par rapport à x_{k+1}:

$$x_{k+1} = x_k + g(x_k, u_k)\,\delta; \quad k = 0, 1, \ldots, n-1, \; x_0 = c. \tag{10-19}$$

Ainsi, l'intégrale (10-16) sera remplacée par la somme

$$J_n(u) = \sum_{k=0}^{n-1} Q(x_k, u_k)\,\delta, \tag{10-20}$$

où l'on entend par u la suite de commandes utilisées

$$u = (u_0, \ldots, u_{n-1}). \tag{10-21}$$

Maintenant, le problème consiste à choisir des commandes u_0, $u_1, \ldots, u_{n-1}$ telles que la somme (10-21) soit minimale.

Dans beaucoup de problèmes de commande, il est rationnel de poser $\delta = 1$. En particulier, cela est commode lorsque le processus est décomposé d'une façon naturelle en étapes isolées, la commande $u(t)$ restant invariable dans les limites de chaque étape. De cette façon, on arrive à la commande à étapes multiples, examinée au chapitre 6, où x_k et u_k signifient l'état de l'objet et la commande adoptée en début de chaque étape. En adoptant, dans ce cas, la notation

$$x_k + g(x_k, u_k) = T(x_k, u_k), \tag{10-22}$$

on écrira l'équation (10-19) sous la forme

$$x_{k+1} = T(x_k, u_k); \quad k = 0, 1, \ldots, n-1; \; x_0 = c, \tag{10-23}$$

qui coïncide avec l'expression (6-52) définissant la transformation de l'état de l'objet au cours d'une étape dans le processus de commande à étapes multiples. Pour $\delta = 1$, la somme (10-20) prend la forme

$$J_n(u) = \sum_{k=0}^{n-1} Q(x_k, u_k), \tag{10-24}$$

qui coïncide avec l'expression du critère de qualité (6-54) dans le processus de commande à étapes multiples.

Dans ce qui suit, la méthode de programmation dynamique sera examinée conformément aux relations (10-23) et (10-24), c.-à-d. par rapport au processus de commande à étapes multiples. Toutefois, la liaison que nous venons d'établir entre les formulations mathématiques de la forme discrète du problème variationnel et le processus de commande à étapes multiples permet également d'étendre les résultats obtenus à la résolution des problèmes variationnels.

b) Relation de récurrence de la méthode de programmation dynamique

L'optimisation de la commande d'un processus à n étapes consiste à trouver une suite de commandes $u_0, u_1, \ldots, u_{n-1}$ telle que le critère de qualité $J_n(u)$ atteigne sa valeur minimale. Cette valeur minimale du critère de qualité de la commande du processus à n étapes dépendra exclusivement de l'état initial x_0 et on peut la désigner par $f_n(x_0)$. Par définition, on a:

$$f_n(x_0) = \min_{u_0} \min_{u_1} \ldots \min_{u_{n-1}} [Q(x_0, u_0) + + Q(x_1, u_1) + \ldots + Q(x_{n-1}, u_{n-1})]. \qquad (10\text{-}25)$$

Remarquons que le premier terme de cette expression $Q(x_0, u_0)$ dépend seulement de la commande u_0, tandis que les autres termes sont fonctions aussi bien de u_0 que des commandes adoptées aux autres étapes. Ainsi, $Q(x_1, u_1)$ dépend de u_1 mais il dépend aussi de u_0, car $x_1 = T(x_0, u_0)$. Il en est de même des autres termes et c'est pourquoi l'expression (10-25) peut s'écrire

$$f_n(x_0) = \min_{u_0} \{Q(x_0, u_0) + + \min_{u_1} \ldots \min_{u_{n-1}} [Q(x_1, u_1) + \ldots + Q(x_{n-1}, u_{n-1})]\}. \qquad (10\text{-}26)$$

Notons encore que l'expression

$$\min_{u_1} \ldots \min_{u_{n-1}} [Q(x_1, u_1) + \ldots + Q(x_{n-1}, u_{n-1})] \qquad (10\text{-}27)$$

représente la valeur minimale du critère de qualité de la commande du processus à $n - 1$ étapes, dont l'état initial est x_1. Conformément à la définition (10-25), cette grandeur peut être notée $f_{n-1}(x_1)$. De cette façon, on obtient:

$$f_n(x_0) = \min_{u_0} [Q(x_0, u_0) + f_{n-1}(x_1)]. \qquad (10\text{-}28)$$

Ces raisonnements peuvent être réitérés pour un processus à $n - 1$ étapes qui commence par l'état initial x_1. Dans ce cas, la valeur mi-

nimale du critère de qualité de la commande sera

$$f_{n-1}(x_1) = \min_{u_1} [Q(x_1, u_1) + f_{n-2}(x_2)]. \qquad (10\text{-}29)$$

Des raisonnements analogues nous amènent à une expression similaire pour un processus à $n - l$ étapes, qui commence par l'état x_l :

$$f_{n-l}(x_l) = \min_{u_l} [Q(x_l, u_l) + f_{n-(l+1)}(x_{l+1})]. \qquad (10\text{-}30)$$

L'équation (10-30), appelée souvent équation de Bellman, représente une relation de récurrence qui permet de déterminer successivement la commande optimale à chaque étape du processus commandé.

L'idée même de l'optimisation de la commande à chaque étape séparément, s'il est difficile d'optimiser tout d'un coup le processus tout entier, n'est pas nouvelle et est souvent utilisée dans la pratique. Mais, avec cela, on oublie souvent que l'optimisation de chaque étape ne signifie pas l'optimisation du processus tout entier. Ainsi, en cédant une pièce aux échecs, on ne peut pas dire que l'on a obtenu un avantage au point de vue d'un coup isolé, mais ce sacrifice peut être avantageux si l'on considère la partie entière. Les frais d'amortissement peuvent paraître désavantageux pour le moment, mais leur utilité pour l'entreprise se fait voir à la longue.

La particularité de la méthode de programmation dynamique consiste dans le fait qu'elle superpose à la simplicité de résolution du problème d'optimisation de la commande à une étape isolée la perspicacité qui permet de tenir compte des conséquences les plus éloignées que peut entraîner l'étape considérée.

Lorsqu'on utilise la méthode de programmation dynamique, le choix de la commande à une étape isolée ne se fait pas suivant les intérêts de cette étape, qui trouvent leur expression dans la minimisation des pertes justement à cette étape, c.-à-d. de la quantité $Q(x_l, u_l)$, mais au point de vue des intérêts du processus tout entier, qui s'expriment par la minimisation des pertes totales $Q(x_l, u_l) + f_{n-(l+1)}(x_{l+1})$ à toutes les étapes suivantes. Il en découle la propriété fondamentale du processus optimal, qui consiste en ce qui suit : *quels que soient l'état initial et la commande initiale, les commandes suivantes doivent être optimales par rapport à l'état qui représente le résultat de l'adoption de la première commande.*

De la propriété fondamentale de la commande optimale il découle que l'optimisation de la commande pour une étape quelconque d'un processus à étapes multiples consiste seulement dans le choix des commandes suivantes. C'est pourquoi il est commode de prendre en considération non pas les étapes déjà passées, mais celles qui restent, pour amener le processus à l'état final. A ce point de vue, il est commode d'écrire l'équation (10-30) sous une autre forme.

Dans l'expression (10-30), la quantité $n - l$ représente le nombre d'étapes qui restent jusqu'à la fin du processus. Désignons cette quantité par k, et désignons, pour simplifier l'écriture, les quantités $x_l = x_{n-k}$ et $u_l = u_{n-k}$ par x et u. Elles vont représenter l'état de l'objet et la commande utilisée à k étapes à compter de la fin du processus. Désignons par x' l'état suivant, c.-à-d. l'état auquel passe l'objet à partir de l'état x en utilisant la commande u. Cela correspond à $x_{n-(l+1)}$ en notations précédentes. De cette façon, l'équation (10-23) s'écrira comme suit

$$x' = T(x, u), \tag{10-31}$$

et la relation de récurrence (10-30) prendra la forme

$$f_k(x) = \min_u [Q(x, u) + f_{k-1}(x')]. \tag{10-32}$$

c) Calculs liés à la programmation dynamique

La programmation dynamique est une méthode numérique de résolution du problème d'optimisation de la commande, donc elle est liée à l'exécution d'un volume important de calculs. Mais cette circonstance perd beaucoup de son importance si l'on tient compte du fait que les calculs correspondants seront effectués à l'aide des calculatrices électroniques.

La détermination de la commande optimale à une étape quelconque k comptée à partir de la fin du processus est réalisée d'après les relations (10-31) et (10-32) qu'il est commode de modifier quelque peu pour faciliter les calculs. Désignons par $F_k(x, u)$ la valeur du critère de qualité de la commande d'un processus à k étapes en présence d'une commande optimale aux $k - 1$ dernières étapes, la commande à l'étape initiale étant arbitraire. Alors, la relation (10-32) peut s'écrire sous la forme

$$F_k(x, u) = Q(x, u) + f_{k-1}(x'); \tag{10-33}$$

$$f_k(x) = \min_u F_k(x, u), \tag{10-34}$$

où x' est trouvé d'après (10-31).

Les variables x et u peuvent prendre soit des ensembles finis de valeurs, soit des valeurs qui varient de façon continue dans certaines gammes. Pour ce dernier cas, rendons discrètes les variables en question en séparant des ensembles finis des valeurs équidistantes dans la mesure du possible. Il s'ensuit que dans les deux cas les variables x et u peuvent être considérées comme des éléments des ensembles finis $X = \{x^{(1)}, \ldots, x^{(m)}\}$ et $U \{u^{(1)}, \ldots, u^{(r)}\}$. Il est commode de calculer à l'avance la valeur de $Q(x, u)$ et de la présenter sous la forme d'un tableau sur le produit direct des ensembles $X \times U$.

Tableau 10-1

Tableau de calcul de la méthode de programmation dynamique

x	u	$x' = T(x, u)$	$Q(x, u)$
$x^{(1)}$	$u^{(1)}$ . . . $u^{(r)}$	$T(x^{(1)}, u^{(1)})$ . . . $T(x^{(1)}, u^{(r)})$	$Q(x^{(1)}, u^{(1)})$ . . . $Q(x^{(1)}, u^{(r)})$
$x^{(2)}$	$u^{(1)}$ . . . $u^{(r)}$	$T(x^{(2)}, u^{(1)})$ . . . $T(x^{(2)}, u^{(r)})$	$Q(x^{(2)}, u^{(1)})$ . . . $Q(x^{(2)}, u^{(r)})$
$x^{(3)}$	. . .	. . .	. . .

x	$f_{k-1}(x')$	$F_k(x, u)$	$f_k(x)$	u^*
$x^{(1)}$	$f_{k-1}[T(x^{(1)}, u^{(1)})]$ . . . $f_{k-1}[T(x^{(1)}, u^{(r)})]$	$F_k(x^{(1)}, u^{(1)})$ . . . $F_k(x^{(1)}, u^{(r)})$	$f_k(x^{(1)})$	u_1^*
$x^{(2)}$	$f_{k-1}[T(x^{(2)}, u^{(1)})]$ . . . $f_{k-1}[T(x^{(2)}, u^{(r)})]$	$F_k(x^{(2)}, u^{(1)})$ . . . $F_k(x^{(2)}, u^{(r)})$	$f_k(x^{(2)})$	u_2^*
$x^{(3)}$	. . .	. . .	. . .	. . .

Ce tableau doit être gardé dans la mémoire de la calculatrice électronique.

Les calculs d'après les formules (10-33) et (10-34) sont effectués en remplissant le tableau 10-1. Les deux premières colonnes contiennent toutes les combinaisons possibles de $x \in X$ et $u \in U$, les autres colonnes étant remplies par voie de calcul des grandeurs respectives. Dans les dernières colonnes on inscrit pour chaque $x \in X$ la valeur minimale de $F_k(x, u)$, c.-à-d. on détermine $f_k(x)$ et la commande optimale u^*.

Des calculs analogues doivent être effectués pour chaque étape du processus à étapes multiples en tenant compte du fait que, pour la détermination de $f_k(x)$, il faut avoir le tableau pour $f_{k-1}(x)$, car ce dernier permet de trouver les valeurs de $f_{k-1}(x')$ lorsqu'on remplit le tableau 10-1. Il s'ensuit que les valeurs de $f_{k-1}(x)$ doivent être calculées avant les valeurs de $f_k(x)$. Donc, le calcul des fonctions $f_k(x)$ doit être commencé à partir de la dernière étape du processus à étapes multiples. Si l'on remarque que $f_0(x) = 0$, comme phase initiale de calcul il faut prendre $k = 1$ pour laquelle

$$F_1(x, u) = Q(x, u); \quad f_1(x) = \min_u Q(x, u). \tag{10-35}$$

Ensuite, les calculs sont effectués de façon habituelle pour $k = 2, 3, \ldots, n$.

Après avoir effectué des calculs et rempli les tableaux pour $f_k(x)$ et u^*, pour $k = 1, \ldots, n$, on peut passer à la recherche de la commande optimale de tout le processus pour les conditions initiales données x_0. D'après le tableau pour $f_n(x)$ on trouve u_0^* correspondant à x_0 donné et l'on calcule $x_1 = T(x_0, u_0^*)$. Ensuite, d'après le tableau pour $f_{n-1}(x)$ on trouve u_1^* correspondant à x_1 trouvé et l'on calcule $x_2 = T(x_1, u_1^*)$ et ainsi de suite. Finalement, on obtient la commande optimale $u^* = (u_0^*, u_1^*, \ldots, u_{n-1}^*)$.

Dans le cas où u est une grandeur continue, $F_k(x^{(i)}, u)$ sera une fonction continue de u. Mais le tableau 10-1 ne donne que les valeurs discrètes $F_k(x^{(i)}, u^{(j)})$, $j = 1, \ldots, r$, de cette fonction, le minimum de $F_k(x^{(i)}, u)$ pouvant correspondre à une valeur intermédiaire de u. Dans ce cas, les valeurs de u^* et de $f_k(x^{(i)})$ doivent être déterminées à l'aide des formules d'interpolation. L'utilisation des formules d'interpolation s'avère nécessaire aussi pour la détermination de la commande optimale u_k^* lorsque les valeurs de $x_k = T(x_{k-1}, u_{k-1})$ seront situées entre les valeurs de x données au tableau 10-1.

Exemple 10-1. Problème de la prise d'altitude d'un avion. Un avion vole à la vitesse v_0 à l'altitude h_0. On demande de porter sa vitesse et son altitude de vol respectivement à v_1 et h_1 en assurant une consommation minimale de combustible.

Représentons sur le plan (v, h) le processus de la variation de v et de h. Rendons discrètes les variables en décomposant chacune des gammes de varia-

tion de v et de h en quatre intervalles. Ainsi, les états discrets de l'objet commandé seront représentés par les mailles de la grille donnée sur la figure 10-2. On considère que dans chaque maille de la grille il ne soit possible d'adopter que deux commandes:

$u_i = 0$: on fait varier seulement la vitesse v;

$u_i = 1$: on fait varier seulement l'altitude h.

De cette façon, l'ensemble des commandes admissibles sera l'ensemble $U = \{0, 1\}$.

Sur la figure 10-2 est représentée une des trajectoires possibles correspondant à la commande $u = (01011001)$. Pour estimer cette trajectoire, il faut

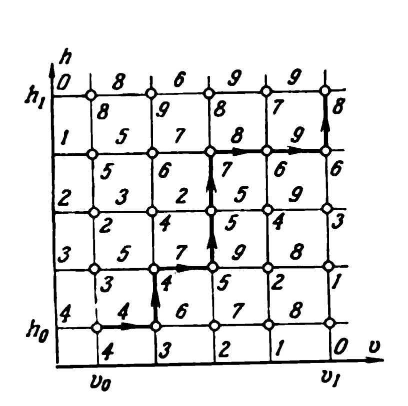


Fig. 10-2. Trajectoire dans le problème de la prise d'altitude

Fig. 10-3. Trajectoire optimale dans le problème de la prise d'altitude

connaître la consommation de combustible sur chaque étape, ce qui nous amène à la fonction objectif $Q(x, u)$. Donnons les valeurs de $Q(x, u)$ sous forme de nombres conventionnels par lesquels est marqué chacun des passages possibles sur la figure 10-2. Pour la trajectoire représentée sur la figure 10-2, la consommation totale de combustible, qui représente la valeur du critère d'efficacité de la commande, est égale à $4 + 4 + 7 + 5 + 7 + 8 + 9 + 8 = 52$.

Pour présenter le processus examiné en tant que processus à étapes multiples, introduisons une méthode adéquate de description des états de l'objet commandé. Désignons les valeurs discrètes de v par les nombres de 0 à 4 à partir de la valeur finale. Procédons de la même façon pour h. Alors, x_{ij} sera l'état caractérisé par $v = i$ et $h = j$ et à partir duquel il reste $i + j$ étapes jusqu'à la fin du processus.

Désignons par X_k l'ensemble des états à partir desquels il reste k étapes jusqu'à la fin du processus. Cet ensemble comprendra tous les états x_{ij} pour lesquels $i + j = k$. En posant $k = 0, 1, 2, \ldots$, on obtient:

$$X_0 = \{x_{00}\}; \quad X_1 = \{x_{10}, x_{01}\}; \quad X_2 = \{x_{20}, x_{11}, x_{02}\}, \text{ etc.}$$

On peut passer maintenant à la résolution du problème. Pour $k = 1$, on a:

$$X_1 = \{x_{10}, x_{01}\}; \quad F_1(x, u) = Q(x, u); \quad f_1(x) = \min_u F_1(x, u).$$

Ces relations servent à la compilation du tableau 10-2 analogue au tableau 10-1.

Tableau 10-2

Calcul de la commande optimale à la dernière étape

x	u	x'	$Q(x, u)$	$f_1(x)$	u^*
x_{10}	0 1	x_{00} —	9 —	9	0
x_{01}	0 1	— x_{00}	— 8	8	1

Pour $k = 2$, on a $X_2 = \{x_{20}, x_{11}, x_{02}\}$;

$$F_2(x, u) = Q(x, u) + f_1(x'), \quad f_2(x) = \min_u F_3(x, u).$$

Les mêmes relations sont utilisées pour former les tableaux 10-3 et 10-4. Des calculs analogues sont effectués pour $k = 3, 4, \ldots$

Pour l'exemple considéré, les données des tableaux 10-2 à 10-4 peuvent être représentées directement sur le plan (v, h) en indiquant les valeurs de $f_k(x)$ sous la forme des nombres inscrits dans les mailles correspondantes de la grille et en représentant u^* par des flèches dirigées vers la maille suivante comme il est visible sur la figure 10-3. Après avoir déterminé les valeurs de $f_k(x)$ et u^* pour toutes les mailles de la grille, on trouve la trajectoire optimale en se déplaçant à partir de la maille initiale suivant les flèches. La commande optimale sera $u^* = (11000110)$ pour laquelle on a $J_n(u^*) = 37$.

Tableau 10-3

Calcul de la commande optimale à l'avant-dernière étape

x	u	x'	$Q(x, u)$	$f_1(x')$	$F_2(x, u)$	$f_2(x)$	u^*
x_{20}	0 1	x_{10} —	9 —	9 —	18 —	18	0
x_{11}	0 1	x_{01} x_{10}	9 7	8 9	17 16	16	1
x_{02}	0 1	— x_{01}	— 6	— 8	— 14	14	1

Tableau 10-4

Calcul de la commande optimale à la troisième étape à compter de la fin du processus

x	u	x'	$Q(x, u)$	$f_2(x')$	$F_3(x, u)$	$f_3(x)$	u^*
x_{30}	0	x_{20}	6	18	24	24	0
	1	—	—	—	—		
x_{21}	0	x_{11}	8	16	24	24	0
	1	x_{20}	8	18	26		
x_{12}	0	x_{02}	9	14	23	22	1
	1	x_{11}	6	16	22		
x_{03}	0	—	—	—	—	17	1
	1	x_{02}	3	14	17		

L'exemple ci-avant met bien en évidence les avantages que présente la méthode de programmation dynamique. Premièrement, cette méthode supprime le problème du minimum absolu et relatif, car la procédure même des calculs montre que l'on trouve toujours le minimum absolu. Deuxièmement, les contraintes de la forme $|u_i| \leqslant \leqslant M_i$, qui constituaient un obstacle sérieux pour l'utilisation des méthodes variationnelles, ne font que simplifier les calculs suivant la méthode de programmation dynamique, car elles réduisent le domaine des commandes admissibles U. Finalement, la méthode de programmation dynamique rend incomparablement plus facile la recherche de la solution optimale par rapport à la méthode du triage des variantes. Un calcul simple permet d'illustrer cette dernière assertion.

Si à chaque étape on a la possibilité d'adopter r commandes différentes, le triage direct exige que chaque commande soit considérée en combinaison avec toutes les commandes possibles aux autres étapes, ce qui donne, pour un processus à n étapes, r^n variantes. Même pour un problème à étapes multiples relativement simple pour $r = = 10$ et $n = 10$, on obtient ainsi un nombre impressionnant de variantes (10^{10}).

En utilisant la méthode de programmation dynamique, lors du choix de la commande à une étape quelconque, pour l'état x, on ne prend pas en considération toutes les continuations possibles, mais seulement celles qui correspondent aux continuations optimales

à partir de l'état $x' = T(x, u)$. Cela permet de ne pas considérer une quantité énorme de variantes privées d'intérêt. Ainsi, si à chaque étape il y a m états possibles et si pour chacun de ces états il y a r commandes, le nombre total de variantes à examiner sera rmn ou r^2n si l'on considère que $r = m$, c.-à-d. que l'on a une commande qui assure le passage de chaque état à n'importe quel nouvel état. Pour $r = 10$ et $n = 10$, cela donne en tout 10^3 variantes.

Mais malgré ses mérites, la méthode de programmation dynamique n'est pas exempte d'inconvénients. Ces inconvénients consistent dans le fait que, pour trouver la commande optimale à partir d'un certain état initial de l'objet, il faut aller de la fin du processus vers son commencement, en déterminant à chaque étape les commandes optimales pour tous les états possibles de l'objet à cette étape, et jusqu'au dernier moment on ignore quelle sera la commande optimale pour l'état initial donné. Finalement on voit que la majeure partie des calculs effectués reste inutilisée, car les résultats concernant les états qui ne se situent pas sur la trajectoire optimale ne sont pas mis en œuvre. A ce point de vue le principe du maximum mentionné plus haut présente des avantages, car il ouvre la voie à la construction d'une trajectoire optimale prise isolément.

d) Commande de l'état final

Au chapitre 6 il a été indiqué que, dans certains problèmes, l'évolution de l'objet durant la réalisation de la commande ne présente pas d'intérêt et ce n'est que l'état x_n, dans lequel passe l'objet à la fin du processus de commande, qui compte. Dans ce cas, en tant que critère de qualité de la commande on prend la valeur de la fonction objectif à la fin du processus de commande, c.-à-d. la quantité

$$q = q(x_n). \tag{10-36}$$

Désignons comme auparavant par x et u l'état de l'objet et la commande utilisée à k étapes de la fin du processus, et par $x' = T(x, u)$ l'état suivant, c.-à-d. l'état de la $(k - 1)$-ième étape à compter de la fin du processus. L'état x_n, donc la fonction objectif $q(x_n)$ aussi, est déterminé aussi bien par la valeur initiale de x que par toutes les commandes qui suivent.

Pour obtenir la relation de récurrence définissant la commande optimale, désignons par $F_k(x, u)$ la valeur de la fonction objectif $q(x_n)$ à k étapes de la fin du processus, pour l'état initial x, pour une commande arbitraire u à l'étape initiale et en présence d'une commande optimale aux $k - 1$ étapes restantes, et par $f_k(x)$ la valeur de $q(x_n)$ pour une commande optimale à toutes les k étapes qui restent jusqu'à la fin du processus, de sorte que

$$f_k(x) = \min_u F_k(x, u). \tag{10-37}$$

La valeur de $F_k(x, u)$ peut être obtenue en raisonnant comme suit. La commande arbitraire u adoptée à l'étape initiale fait passer l'objet commandé de l'état x à l'état $x' = T(x, u)$. Pour obtenir $q(x_n) = F_k(x, u)$, il faut adopter la commande optimale aux $k-1$ étapes qui restent, ce qui donne $q(x_n) = f_{k-1}(x') = f_{k-1}[T(x, u)]$. Or, c'est la valeur de $F_k(x, u)$. De cette façon,

$$F_k(x, u) = f_{k-1}[T(x, u)]. \quad (10\text{-}38)$$

En portant les valeurs de $F_k(x, u)$ dans (10-37), on obtient:

$$f_k(x) = \min_u f_{k-1}[T(x, u)]. \quad (10\text{-}39)$$

La formule (10-39) est parfaitement analogue, d'après sa structure, à la formule (10-32). Si l'on lui applique la méthode de calcul déjà connue et si l'on prend en considération que

$$f_0(x) = q(x), \quad (10\text{-}40)$$

on peut trouver la commande optimale pour tout état initial du processus à étapes multiples.

e) Relation de récurrence pour les processus markoviens

Examinons l'application de la méthode de programmation dynamique à l'optimisation des processus stochastiques dans le cas où le processus stochastique est une chaîne de Markov commandée.

Au paragraphe 5-6 on a vu qu'une chaîne de Markov peut être donnée sous la forme d'une matrice de probabilités des passages $\mathbf{P} = \| p_{ij} \|$ d'ordre $L \times L$. Pour que la chaîne de Markov soit commandée, il faut que l'on puisse changer les probabilités des passages p_{ij} en agissant de l'extérieur. Supposons que l'on puisse conduire le processus markovien par N procédés différents et qu'au k-ième procédé corresponde la matrice de probabilités des passages $P_k = \| p_{ij}^{(k)} \|$, $k = 1, \ldots, N$. Chaque procédé de conduire le processus markovien sera appelé stratégie.

En outre, nous allons considérer que l'on a la possibilité d'estimer chaque procédé de réalisation du processus en se donnant la matrice de paiement ou, ce qui est plus commode pour le cas considéré, la matrice de gains $\mathbf{R}_k = \| r_{ij}^{(k)} \|$. Les grandeurs $r_{ij}^{(k)}$, $i, j = 1, \ldots, L$; $k = 1, \ldots, N$ expriment le gain d'une étape au passage du processus de l'état i à l'état j en cas d'adoption de la k-ième stratégie.

Exemple 10-2. La fabrique de téléviseurs de l'exemple 5-9 se trouvant dans l'état 1 peut augmenter la demande à l'aide de la réclame, mais celle-ci nécessite des frais supplémentaires qui font diminuer les gains. Si la fabrique se trouve dans l'état 2, elle peut accélérer le passage à l'état 1 en augmentant les frais d'études. Dans ce cas, la stratégie 1 consiste à ne pas subir des frais de réclame

et d'études, tandis que la stratégie 2 prévoit justement ces frais. Les matrices de probabilités des passages et les matrices de gains pour ces deux stratégies peuvent être les suivantes:

$$\mathbf{P}_1 = \begin{Vmatrix} 0,5 & 0,5 \\ 0,4 & 0,6 \end{Vmatrix}; \qquad \mathbf{R}_1 = \begin{Vmatrix} 9 & 3 \\ 3 & -7 \end{Vmatrix};$$

$$\mathbf{P}_2 = \begin{Vmatrix} 0,8 & 0,2 \\ 0,7 & 0,3 \end{Vmatrix}; \qquad \mathbf{R}_2 = \begin{Vmatrix} 4 & 4 \\ 1 & -19 \end{Vmatrix}.$$

On se propose de trouver la méthode de choix de la stratégie qui assure les gains maximaux pendant n étapes du processus markovien commandé.

Désignons par $f_i(n)$ la valeur maximale des gains réalisés en n étapes à partir de l'état i. Supposons qu'en une étape on passe de l'état i à l'état j en réalisant le gain $r_{ij}^{(k)}$, les $n-1$ passages qui restent étant réalisés de la façon optimale. Ainsi, les gains totaux réalisés pendant n étapes seront $r_{ij}^{(k)} + f_j(n-1)$.

Mais en réalité le premier passage de la forme (i, j) est réalisé avec la probabilité $p_{ij}^{(k)}$. Etant donné que le premier passage est aléatoire, il faut prendre en considération la possibilité de passage à tous les états possibles $j \in \{1, \ldots, N\}$. C'est pourquoi les gains espérés seront:

$$F_i(n, k) = \sum_{j=1}^{N} p_{ij}^{(k)} [r_{ij}^{(k)} + f_j(n-1)] = q_i^{(k)} + \sum_{j=1}^{N} p_{ij}^{(k)} f_j(n-1), \quad (10\text{-}41)$$

où la quantité

$$q_i^{(k)} = \sum_{j=1}^{N} p_{ij}^{(k)} r_{ij}^{(k)} \quad (10\text{-}42)$$

donne les gains réalisés durant une étape à partir de l'état i et en utilisant la stratégie k.

L'expression (10-41) contient la stratégie k adoptée à l'étape initiale et dont le choix peut assurer les gains maximaux espérés $f_i(n)$, ce qui nous amène à la relation de récurrence pour le choix de la stratégie k à la première étape d'un processus à n étapes:

$$f_i(n) = \max_k [q_i^{(k)} + \sum_{j=1} p_{ij}^{(k)} f_j(n-1)]. \quad (10\text{-}43)$$

Il est commode d'effectuer les calculs d'après la formule (10-43) en utilisant un tableau analogue à celui examiné aux paragraphes précédents. Ci-après sont donnés les tableaux 10-5 et 10-6 pour le problème de l'exemple 10-2 correspondant respectivement à $n = 1$ et $n = 2$.

Tableau 10-5

$n = 1$

i	k	$p_{ij}^{(k)}$		$r_{ij}^{(k)}$		$q_i^{(k)}$	$f_i(1)$	k^*
		$j=1$	$j=2$	$j=1$	$j=2$			
1	1	0,5	0,5	9	3	6	6	1
	2	0,8	0,2	4	4	4		
2	1	0,4	0,6	3	—7	—3	—3	1
	2	0,7	0,3	1	—19	—5	.	

Tableau 10-6

$n = 2$

i	k	$q_i^{(k)}$	$p_{ij}^{(k)}$		$p_{ij}^{(k)} f_j(1)$		$F_i(2, k)$	$f_i(2)$	k^*
			$j=1$	$j=2$	$j=1$	$j=2$			
1	1	6	0,5	0,5	3	—1,5	7,5	8,2	2
	2	4	0,8	0,2	4,8	—0,6	8,2		
2	1	—3	0,4	0,6	2,4	—1,8	—2,4	—1,7	2
	2	—5	0,7	0,3	4,2	—0,9	—1,7		

Pour $n > 2$, les tableaux sont du type du tableau pour $n = 2$.

PROBLÈMES AU CHAPITRE 10

10-1. Continuer la résolution de l'exemple 10-1 et calculer la trajectoire optimale tout entière.

10-2. En considérant la grille de la figure 10-3 en tant que graphe, trouver la trajectoire optimale à l'aide des méthodes de détermination du plus court chemin dans le graphe, exposées au paragraphe 2-2. Comparer cette solution à celle de la méthode de programmation dynamique.

10-3. Trouver la commande optimale pour un processus de commande à trois étapes de la grandeur scalaire x. En début du processus, x peut être un entier quelconque compris entre —10 et +10. Aux différentes étapes, on réalise des transformations de la forme $x' = T(x, u) = x + u$, les valeurs admissibles de u étant définies par les ensembles:

$\{-1, 0, +1\}$ à la première étape;

$\{-4, 0, +4\}$ à la deuxième étape;

$\{-9, 0, +9\}$ à la troisième étape.

Le but du processus est de porter en trois étapes la grandeur x à une valeur donnée prise égale à zéro. Les pertes sont estimées par le carré de l'écart de x par rapport à 0 en fin de processus, c.-à-d. par la quantité $q(x) = (x - 0)^2 = x^2$.

BIBLIOGRAPHIE

1. Wiener N. Cybernetics or control and communication in the animal and the machine.
2. Философия естествознания [Philisophie des sciences naturelles]. M., Изд-во АН СССР, 1966, вып. 1.
3. Beer S. Cybernetics and management.
4. Ashby W. R. An introduction to cybernetics. London, Chapman and Hall, 1956.
5. Кузин Л. Т. Основы кибернетики [Kousine L. Eléments de cybernétique]. M., Изд. МИФИ, 1970.
6. Булгаков А. А. Электронные устройства автоматического управления [Boulgakov A. Equipements électroniques de commande automatique]. М.-Л., Госэнергоиздат, 1958.
7. Kemeny J. G., Snell J. L., Thompson G. L. Introduction to finite mathematics.
8. Шиханович Ю. А. Введение в современную математику [Chikhanovitch Y. Introduction aux mathématiques modernes]. M., « Наука », 1965.
9. Kleene St. C. Introduction to metamathematics. New York-Toronto, D. Van Nostrand Co., 1952.
10. Carr Ch. R., Howe Ch. W. Quantitative decision procedures in management and economics.
11. Félix L. Exposé moderne des mathématiques élémentaires. P., Dunod, 1959.
12. Berge Cl. Théorie des graphes et ses applications. P., Dunod, 1958.
13. Ore O. Theory of graphs.
14. Ore O. Graphs and their uses.
15. Ford L. R., Felkerson D. R. Flows in networks.
16. Kolmogorov A., Fomine S. Eléments de la théorie des fonctions et de l'analyse fonctionnelle. Editions Mir, Moscou, 1974.
17. Шрейдер Ю. А. Что такое расстояние? [Chreïder Y. Qu'est-ce que la distance?]. M. Физматгиз, 1963.
18. Куликовский Р. Оптимальные и адаптивные процессы в системах автоматического регулирования [Koulikovski R. Processus optimaux et adaptatifs dans les systèmes de régulation automatique]. M., «Наука», 1967.
19. Харкевич А. А. Борьба с помехами [Kharkévitch A. Antiparasitage radio-électrique]. M., « Наука », 1965.
20. Березин И. С., Жидков Н. П. Методы вычислений [Bérésine N., Jidkov N. Méthodes de calcul]. Т. 1. M., Физматгиз, 1959.
21. Sebestyen G. S. Decision-making processes in pattern recognition.
22. Вопросы статистической теории распознавания. M., « Советское радио », 1967. Авт. : Барабаш Ю. Л., Варский Б. В. и др. [Problèmes de la théorie statistique de l'identification. Par Barabach A., Varski B. et autres].
23. Новиков П. С. Элементы математической логики [Novikov P. Eléments de logique mathématique]. M., Физматгиз, 1959.
24. Глушков В. М. Введение в кибернетику [Glouchkov V. Introduction à la cybernétique]. Киев, Изд-во АН УССР, 1964.

25. Логика, автоматы, алгоритмы. М., Физматгиз, 1963. Авт. : Айзерман М. А., Гусев Л. А. и др. [Logique, automates, algorithmes. Par Aizerman M., Goussev L. et autres].
26. Berkeley Ed. S. Symbolic logic and intelligent machines. New York, Reinhold publ. corporation ; London, Chapman and Hall.
27. Arbib M. A. Brains, machines and mathematics.
28. Caldwell S. H. Switching circuits and logical design. New York, Wiley, 1958.
29. Richards R. K. Digital computer components and circuits. Toronto a.o., Van Nostrand co., 1958.
30. Richards R. K. Arithmetic operation in digital computers. Torondo a.o., Van Nostrand co., 1955.
31. Глушков В. М. Синтез цифровых автоматов [Glouchkov V. Synthèse d'automates numériques]. М., Физматгиз, 1962.
32. Automata studies. Ed. by C. E. Shannon and J. McCarthy. Princeton-New Jersey, Princeton univ. press, 1956.
33. Ventsel H. Théorie des probabilités. Editions Mir., Moscou, 1973.
34. Гнеденко Б. В. Курс теории вероятностей [Ghédenko B. Cours de la théorie des probabilités]. М., « Наука », 1965.
35. Wilks S. St. Mathematical statistics. Princeton, New Jersey, Princeton university press, 1946.
36. Смирнов Н. В., Дунин-Барковский И. В. Курс теории вероятностей и математической статистики [Smirnov N., Dounine-Barkovski I. Cours de la théorie des probabilités et de la statistique mathématique]. М., « Наука », 1965.
37. Gnédenko B., Béliaev Y., Soloviev A. Méthodes mathématiques en théorie de la fiabilité. Editions Mir, Moscou, 1972.
38. Lehmann E. L. Testing statistical hypotheses. New York, Wiley ; London, Chapman and Hall.
39. Основы автоматического управления. Под ред. В. С. Пугачева [Eléments de commande automatique. Rédigé par Pougatchev V.]. Физматгиз, 1963.
40. Вентцель Е. С. Исследование операций [Ventsel H. Recherche opérationnelle]. М., « Советское радио », 1972.
41. Churchman C. W., Ackoff R. L., Arnoff E. L. Introduction to operations research.
42. Болтянский В. Г. Математика и оптимальное управление [Boltianski V. Les mathématiques et la commande optimale]. М., « Знание », 1968.
43. Цыпкин Я. З. Адаптация и обучение в автоматических системах [Tsipkine Y. Adaptation et enseignement dans les systèmes automatiques]. М., « Наука », 1968.
44. Hadley G. Non linear and dynamic programming.
45. Математические модели и методы оптимального планирования [Modèles mathématiques et méthodes de planification optimale].— Труды Института математики СО АН СССР. Новосибирск, « Наука », 1966.
46. Корбут А. А., Финкельштейн Ю. Д. Дискретное программирование [Korboute A., Finkelstein Y. Programmation discrète]. М., « Наука », 1969.
47. Issacs R., Differential games.
48. Бронштейн И. Н., Семендяев К. А. Справочник по математике [Bronstein I., Sémendiaev K. Aide-mémoire de mathématiques]. М., « Наука », 1964.
49. Démidovitch B., Marone I. Eléments de calcul numérique. Editions Mir, Moscou, 1973.
50. Барсов А. С. Что такое линейное программирование? [Barsov A. Qu'est-ce que la programmation linéaire?]. М., Физматгиз, 1959.
51. Карпелевич Ф. И., Садовский Л. Е. Элементы линейной алгебры и линейного программирования [Karpélévitch F., Sadovski L. Eléments d'algèbre linéaire et de programmation linéaire]. М., Физматгиз, 1963.

52. Юдин Д. Б., Гольштейн Е. Г. Задачи и методы линейного программирования [Youdine D., Golstein E. Problèmes et méthodes de programmation linéaire]. M., « Советское радио », 1964.
53. Dantzig G. B. Linear programming and extensions.
54. Blackwell D., Girshick M. A. Theory of games and statistical decisions. New York, John and Sons, inc. London, Chapman and Hall.
55. McKinsey J. C. C. Introduction to the theory of games. 1-th ed. New York, a.o., McGraw-Hill, 1952.
56. Von Neumann J., Morgenstern O. Theory of games and economic behavior. Princeton univ. press., 1947.
57. Чернов Г., Мозес Л. Элементарная теория статистических решений [Tchernov G., Mosess L. Théorie élémentaire des décisions statistiques]. M., « Советское радио », 1962.
58. Middleton D. An introduction to statistical communication theory, 1960.
59. Vald A. Sequential analysis. New York, John Wiley and Sons, inc. London, Champan and Hall.
60. Башаринов А. Е., Флейшман Б. С. Методы статистического последовательного анализа и их приложения [Bacharinov A., Fleichman B. Méthodes d'analyse statistique séquenticlle et leurs applications]. M., « Советское радио », 1962.
61. Эльсгольц Л. Э. Дифференциальные уравнения и вариационное исчисление [Elsgolts L. Equations différentielles et calcul des variations]. M., « Наука », 1965.
62. Pontriaguine L., Boltianski V., Gamkrélidzé R., Michtchenko E. Théorie mathématique des processus optimaux. Editions Mir, Moscou, 1974.
63. Болтянский В. Г. Математические методы оптимального управления [Boltianski V. Méthodes mahématiques de la commande optimale]. M., « Наука », 1966.
64. Розоноэр Л. И. Принцип максимума Л. С. Понтрягина в теории оптимальных систем. [Rosonoer L. Principe du maximum de L. Pontriaguine dans la théorie des systèmes optimaux]. — « Автоматика и телемеханика », 1959, т. 20, № 10-12.
65. Bellman R., Dynamic Programming. New Jersey, Princeton univ. press, N. Y., 1957.
66. Bellman R. Adaptive control processes : a guide tour.
67. Вентцель Е. С. Элементы динамического программирования [Ventsel H. Eléments de programmation dynamique]. M., « Наука », 1964.

INDEX

TABLE DES MATIÈRES

Deuxième partie

OPTIMISATION DES PROCESSUS DE COMMANDE